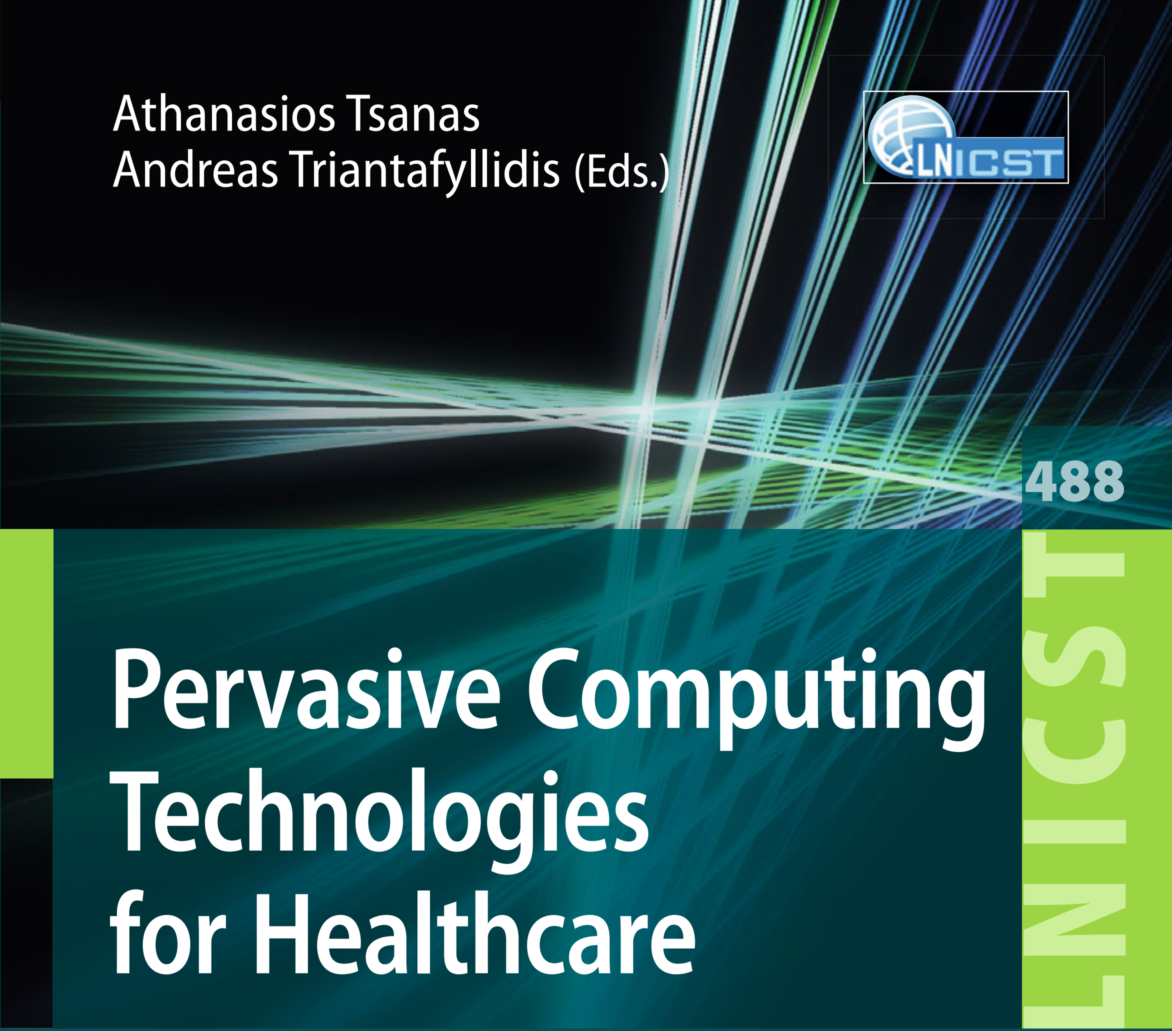


Athanasios Tsanas
Andreas Triantafyllidis (Eds.)

LNICST

488

Pervasive Computing Technologies for Healthcare

16th EAI International Conference, PervasiveHealth 2022
Thessaloniki, Greece, December 12–14, 2022
Proceedings

EAI

Springer

Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering

488

Editorial Board Members

Ozgur Akan, *Middle East Technical University, Ankara, Türkiye*
Paolo Bellavista, *University of Bologna, Bologna, Italy*
Jiannong Cao, *Hong Kong Polytechnic University, Hong Kong, China*
Geoffrey Coulson, *Lancaster University, Lancaster, UK*
Falko Dressler, *University of Erlangen, Erlangen, Germany*
Domenico Ferrari, *Università Cattolica Piacenza, Piacenza, Italy*
Mario Gerla, *UCLA, Los Angeles, USA*
Hisashi Kobayashi, *Princeton University, Princeton, USA*
Sergio Palazzo, *University of Catania, Catania, Italy*
Sartaj Sahni, *University of Florida, Gainesville, USA*
Xuemin Shen, *University of Waterloo, Waterloo, Canada*
Mircea Stan, *University of Virginia, Charlottesville, USA*
Xiaohua Jia, *City University of Hong Kong, Kowloon, Hong Kong*
Albert Y. Zomaya, *University of Sydney, Sydney, Australia*

The LNICST series publishes ICST's conferences, symposia and workshops.

LNICST reports state-of-the-art results in areas related to the scope of the Institute.

The type of material published includes

- Proceedings (published in time for the respective event)
- Other edited monographs (such as project reports or invited volumes)

LNICST topics span the following areas:

- General Computer Science
- E-Economy
- E-Medicine
- Knowledge Management
- Multimedia
- Operations, Management and Policy
- Social Informatics
- Systems

Athanasios Tsanas · Andreas Triantafyllidis
Editors

Pervasive Computing Technologies for Healthcare

16th EAI International Conference, PervasiveHealth 2022
Thessaloniki, Greece, December 12–14, 2022
Proceedings

Springer

Editors
Athanasios Tsanas
University of Edinburgh
Edinburgh, UK

Andreas Triantafyllidis
Centre for Research and Technology Hellas
Thessaloniki, Greece

ISSN 1867-8211 ISSN 1867-822X (electronic)
Lecture Notes of the Institute for Computer Sciences, Social Informatics
and Telecommunications Engineering
ISBN 978-3-031-34585-2 ISBN 978-3-031-34586-9 (eBook)
https://doi.org/10.1007/978-3-031-34586-9

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

We are delighted to introduce to the scientific community the proceedings of the 16th European Alliance for Innovation (EAI) International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2022), which was held on 12–13 December 2022 in Thessaloniki, Greece. The conference brought together researchers, engineers and practitioners around the world focusing on the design, implementation and evaluation of ubiquitous computing tools and services for healthcare and wellbeing. The PervasiveHealth conference this year focused on telemedicine, new technologies designed and developed to face the challenges of COVID-19, and the way healthcare systems should be re-designed building on close collaboration between different stakeholders including partners from industry, academia and decision-makers to create advanced healthcare systems.

The technical program of PervasiveHealth 2022 consisted of 45 full papers carefully selected following rigorous double-blind peer review out of 120 submitted papers that were sent for comments by expert reviewers. There were additional submissions which did not fit into the remit of the conference or fell short of the initial editorial screening and hence were not peer-reviewed; this highlights that PervasiveHealth 2022 was a highly selective forum with a less than 35% acceptance rate. We organized the conference in 9 broad thematic sessions, which were: Session 1: Personal Informatics and Wearable Devices; Session 2: Computer Vision; Session 3: Pervasive Health for COVID-19; Session 4: Machine Learning, Human Activity Recognition and Speech Recognition; Session 5: Software Frameworks and Interoperability; Session 6: Facial Recognition, Gesture Recognition and Object Detection; Session 7: Machine Learning, Predictive Models and Personalized Healthcare; Session 8: Human-Centred Design of Pervasive Health Solutions; Session 9: Personalized Healthcare. In addition to the high-quality technical paper presentations, the technical program also featured one keynote speech, one technical workshop and one tutorial. The keynote speaker was Leontios Hadjileontiadis from the Aristotle University of Thessaloniki, Greece, and Khalifa University, United Arab Emirates, presenting his latest research with the title "*Towards Digital Phenotyping: New AI-based Digital Biomarkers*". The workshop organized was entitled "*IoT-HR: Workshop on Internet of Things in Health Research*", and was chaired by Dario Salvi from Malmö University, Sweden, and Francesco Potortì from the National Research Council of Italy. The workshop aimed to explore the challenges of solutions based on the Internet of Things for healthcare together with researchers and practitioners from academia and industry. Finally, the tutorial entitled "*Wearable Intelligence for Parkinson Disease Diagnosis in Free-living Environment: A Huawei Smartwatch System Case Study*", focused on wearable intelligence and was organized by Xulong Wang, Peng Yue, and Po Yang, from the University of Sheffield, UK.

We sincerely appreciate the strong support and guidance from EAI and the steering committee of the conference. It was also a great pleasure to work with such an excellent organizing committee team. We would like to thank in this regard (in no particular order): Konstantinos Votis, Pedro Gomez-Vilda, Dimitrios Fotiadis, Manolis Tsiknakis, Haridimos Kondylakis, Sofia Segkouli, Siddharth Arora, Dario Salvi, Nicos Maglaveras, and Kathrin Cresswell, for their support and hard work in organizing the conference. We would like also to thank the Chairman of the Centre for Research and Technology Hellas (CERTH), Dimitrios Tzovaras, for his unwavering support during the conference, and all personnel at CERTH, who demonstrated the utmost commitment towards all organizational aspects. We are also grateful to the reviewers, who provided timely and constructive feedback to our authors. Last but not least, we would like to express our gratitude to all the enthusiastic presenters and participants in the conference for their thought-provoking contributions and engagement at the event.

Athanasios Tsanas
Andreas Triantafyllidis

Organization

Steering Committee

Imrich Chlamtac	University of Trento, Italy

Organizing Committee

General Chair

Athanasios Tsanas	University of Edinburgh, UK

General Co-Chair

Andreas Triantafyllidis	Centre for Research and Technology Hellas, Greece

Technical Program Committee Chairs

Pedro Gomez-Vilda	Universidad Politécnica de Madrid, Spain
Dimitrios Fotiadis	University of Ioannina, Greece

Technical Program Committee Co-Chair

Konstantinos Votis	Centre for Research and Technology Hellas, Greece

Web Chair

Dario Salvi	Malmö University, Sweden

Publicity and Social Media Chair

Sofia Segkouli	Centre for Research and Technology Hellas, Greece

Workshops Chair

Haridimos Kondylakis	ICS FORTH, Greece

Publications Chair

Manolis Tsiknakis	ICS FORTH, Greece

Tutorials Chair

Siddharth Arora	University of Oxford, UK

Demos Chair

Kathrin Cresswell	University of Edinburgh, UK

Local Chair

Nicos Maglaveras	Aristotle University of Thessaloniki, Greece

Technical Program Committee

Paolo Barsocchi	ISTI, Italy
Chris Paton	University of Oxford, UK
Carmelo Velardo	Sensyne Health, UK
Siddharth Arora	University of Oxford, UK
Karim Lekadir	University of Barcelona, Spain
Manolis Tsiknakis	ICS FORTH, Greece
Evangelia Zacharaki	University of Patras, Greece
Haridimos Kondylakis	ICS FORTH, Greece
Giuseppe Fico	Universidad Politécnica de Madrid
Georgios Theodorakopoulos	Cardiff University, UK
Haodi Zhong	King’s College London, UK
Honghan Wu	UCL, UK
Holly Tibble	University of Edinburgh,UK
Tracey Chantler	The London School of Hygiene & Tropical Medicine, UK
Francisco Lupiáñez-Villanueva	Open Evidence, Spain
Asimina Kiourti	Ohio State University, USA
Ioannis Katakis	University of Nicosia, Cyprus

Gaetano Valenza	University of Pisa, Italy
Iosif Mporas	University of Hertfordshire, UK
Ahmar Shah	University of Edinburgh, UK
Stelios Zygouris	Centre for Research and Technology Hellas, Greece
Eirini Lithoxoidou	Centre for Research and Technology Hellas, Greece
Anastasios Alexiadis	Centre for Research and Technology Hellas, Greece
Sofia Segkouli	Centre for Research and Technology Hellas, Greece
Ilias Kalamaras	Centre for Research and Technology Hellas, Greece
Antonios Lalas	Centre for Research and Technology Hellas, Greece
Varvara Kalokyri	ICS FORTH, Greece
Lefteris Koumakis	ICS FORTH, Greece
Alexandros Kanterakis	ICS FORTH, Greece
Eleni Kolokotroni	NTUA, Greece
Georgios Manikis	Karolinska Institute, Sweden
Grigorios Loukides	King’s College London, UK
Honghan Wu	UCL, UK
Gerasimos Arvanitis	University of Patras, Greece
Stavros Nousias	Industrial Systems Institute, Athena Research Center, Greece
Nikos Fazakis	University of Patras, Greece
Ioannis Konstantoulas	University of Patras, Greece
Thomas Papastergiou	University of Montpellier, France
Riccardo Colella	University of Salento, Italy

Contents

Personal Informatics and Wearable Devices

Robust Respiration Sensing Based on Wi-Fi Beamforming 3
Wenchao Song, Zhu Wang, Zhuo Sun, Hualei Zhang, Bin Guo, Zhiwen Yu, Chih-Chun Ho, and Liming Chen

Heart Rate During Sleep Measured Using Finger-, Wrist- and Chest-Worn Devices: A Comparison Study 18
Nouran Abdalazim, Joseba Aitzol Arbilla Larraza, Leonardo Alchieri, Lidia Alecci, Silvia Santini, and Shkurta Gashi

A Low-Cost Wearable System to Support Upper Limb Rehabilitation in Resource-Constrained Settings 33
Md. Sabbir Ahmed, Shajnush Amir, Samuelson Atiba, Rahat Jahangir Rony, Nervo Verdezoto Dias, Valerie Sparkes, Katarzyna Stawarz, and Nova Ahmed

Computer Vision

Multiclass Semantic Segmentation of Mediterranean Food Images 49
Fotios S. Konstantakopoulos, Eleni I. Georga, and Dimitrios I. Fotiadis

Comparative Study of Machine Learning Methods on Spectroscopy Images for Blood Glucose Estimation 60
Tahsin Kazi, Kiran Ponakaladinne, Maria Valero, Liang Zhao, Hossain Shahriar, and Katherine H. Ingram

Classification of Kidney Tumor Grading on Preoperative Computed Tomography Scans 75
Maryamalsadat Mahootiha, Hemin Ali Qadir, Jacob Bergsland, and Ilangko Balasingham

IoT-HR: Internet of Things in Health Research

An IoT-Based System for the Study of Neuropathic Pain in Spinal Cord Injury 93
Dario Salvi, Gent Ymeri, Daniel Jimeno, Vanesa Soto-León, Yolanda Pérez-Borrego, Carl Magnus Olsson, and Carmen Carrasco-Lopez

IoT Smart Shoe Solution for Neuromuscular Disease Monitoring 104
Davide La Rosa, Filippo Palumbo, Alessandra Ronca, Francesco Sansone, Mario Tesconi, Alessandro Tonacci, and Raffaele Conte

Compliance and Usability of an Asthma Home Monitoring System 116
Kevin C. H. Tsang, Hilary Pinnock, Andrew M. Wilson, Dario Salvi, Carl Magnus Olsson, and Syed Ahmar Shah

Assessing Older Adult's Gait Speed with Wearable Accelerometers in Community Settings: Validity and Reliability Study 127
Antonio Cobo, Elena Villalba-Mora, Rodrigo Pérez-Rodríguez, Juan E. Medina, Paula Robles-Mateos, Ángel Rodríguez-Laso, and Leocadio Rodríguez-Mañas

Healthcare Based on IoT Using Arduino, MPU6050 Accelerometer and Gyroscope Sensor and FSR-174 Strain Gauge for Fatigue 138
Sara Fernández-Canosa, Verónica López González, Antonio Consoli, and Vanesa Soto-León

Smartphone-Based Strategy for Quality-of-Life Monitoring in Head and Neck Cancer Survivors 149
Laura Lopez-Perez, Itziar Alonso, Elizabeth Filippodou, Franco Mercalli, Stefano Cavalieri, Elena Martinelli, Lisa Licitra, Anastassios Manos, María Fernanda Cabrera-Umpierrez, and Giuseppe Fico

Pervasive Health for COVID-19

COVID-19 Classification Algorithm Based on Privacy Preserving Federated Learning 161
Changqing Ji, Cheng Baoluo, Gao Zhiyong, Qin Jing, and Wang Zumin

Flattening the Curve Through Reinforcement Learning Driven Test and Trace Policies 174
Andrei C. Rusu, Katayoun Farrahi, and Mahesan Niranjan

Connecting Self-reported COVID-19 Needs with Social Determinants of Health 207
Jessica A. Pater, Tammy Toscos, Mindy Flanagan, Michelle Drouin, Deborah McMahan, Meg Distler, Patti Hayes, and Nelson Peters

Machine Learning, Human Activity Recognition and Speech Recognition

Up-Sampling Active Learning: An Activity Recognition Method for Parkinson's Disease Patients 229
Peng Yue, Xiang Wang, Yu Yang, Jun Qi, and Po Yang

Your Day in Your Pocket: Complex Activity Recognition from Smartphone Accelerometers 247
Emma Bouton-Bessac, Lakmal Meegahapola, and Daniel Gatica-Perez

Research on Passive Assessment of Parkinson's Disease Utilising Speech Biomarkers 259
Daniel Kovac, Jiri Mekyska, Lubos Brabenec, Milena Kostalova, and Irena Rektorova

Millimeter Wave Radar Sensing Technology for Filipino Sign Language Recognition 274
Jorelle Aaron Herrera, Almira Astrid Muro, Philip Luis Tuason III, Paul Vincent Alpano, and Jhoanna Rhodette Pedrasa

Dehydration Scan: An Artificial Intelligence Assisted Smartphone-Based System for Early Detection of Dehydration 289
Priyeta Saha, Syed Muhammad Ibne Zulfiker, Tanzima Hashem, and Khandker Aftarul Islam

Software Frameworks and Interoperability

Experiencer: An Open-Source Context-Sensitive Wearable Experience Sampling Tool 315
Alireza Khanshan, Pieter Van Gorp, and Panos Markopoulos

BONVITA: Enabling Integrated Self-Care for Improving Chronic Patient's Wellbeing 332
Haridimos Kondylakis, Angelina Kouroubali, and Dimitrios G. Katehakis

Data Analytics for Health and Connected Care: Ontology, Knowledge Graph and Applications 344
Bram Steenwinckel, Mathias De Brouwer, Marija Stojchevska, Jeroen Van Der Donckt, Jelle Nelis, Joeri Ruyssinck, Joachim van der Herten, Koen Casier, Jan Van Ooteghem, Pieter Crombez, Filip De Turck, Sofie Van Hoecke, and Femke Ongenae

Health Data Semantics: Exploring Requirements for Sustainable Health Systems 361
Mate Bestek and Erik Grönvall

Facial Recognition, Gesture Recognition and Object Detection

A Quantitative Comparison of Manual vs. Automated Facial Coding Using Real Life Observations of Fathers 379
Romana Burgess, Iryna Culpin, Helen Bould, Rebecca Pearson, and Ian Nabney

Knowledge-Driven Dialogue and Visual Perception for Smart Orofacial Rehabilitation 397
Jacobo López-Fernández, Luis Unzueta, Meritxell Garcia, Maia Aguirre, Ariane Méndez, and Arantza del Pozo

MM4Drone: A Multi-spectral Image and mmWave Radar Approach for Identifying Mosquito Breeding Grounds via Aerial Drones 412
K. T. Y. Mahima, Malith Weerasekara, Kasun De Zoysa, Chamath Keppitiyagama, Markus Flierl, Luca Mottola, and Thiemo Voigt

Machine Learning, Predictive Models and Personalised Healthcare

Spatio-Temporal Predictive Modeling for Placement of Substance Use Disorder Treatment Facilities in the Midwestern U.S 429
Jessica A. Pater, Shion Guha, Rachel Pfafman, Connie Kerrigan, and Tammy Toscos

Linking Data Collected from Mobile Phones with Symptoms Level in Parkinson's Disease: Data Exploration of the mPower Study 439
Gent Ymeri, Dario Salvi, and Carl Magnus Olsson

An Exploratory Study of the Value of Vital Signs on the Short-Term Prediction of Subcutaneous Glucose Concentration in Type 1 Diabetes – The GlucoseML Study 450
Daphne N. Katsarou, Eleni I. Georga, Maria Christou, Stelios Tigas, Costas Papaloukas, and Dimitrios I. Fotiadis

Less is More: Leveraging Digital Behavioral Markers for Real-Time Identification of Loneliness in Resource-Limited Settings 460
Md. Sabbir Ahmed and Nova Ahmed

Toward Understanding Users' Interactions with a Mental Health App: An Association Rule Mining Approach 477
Alaa Alslaity, Gerry Chan, Richard Wilson, and Rita Orji

Application of Shapley Additive Explanation Towards Determining Personalized Triage from Health Checkup Data 496
Luo Sixian, Yosuke Imamura, and Ashir Ahmed

Human-Centred Design of Pervasive Health Solutions

Dance Mat Fun - A Participatory Design of Exergames for Children with Disabilities 513
Gonçalo Roxo, Rui Nóbrega, and Rui Neves Madeira

Designing Hearing Aids to Mitigate Perceived Stigma Associated with Hearing Impairment 528
Guanhua Sun, Leila Aflatoony, and Wendell Wilson

Demands on User Interfaces for People with Intellectual Disabilities, Their Requirements, and Adjustments 540
Melinda C. Braun and Matthias Wölfel

A Rule Mining and Bayesian Network Analysis to Explore the Link Between Depression and Digital Behavioral Markers of Games App Usage 557
Md. Sabbir Ahmed, Tanvir Hasan, Md. Mahfuzur Rahman, and Nova Ahmed

mHealth for Medication and Side Effect Monitoring: Patients' Attitudes Toward Smart Devices for Managing Oral Chemotherapy During Lung Cancer Treatment 570
Anna N. Baglione, Sarah Tolman, Chloe Dapaah, Danielle Johnson, Kristen J. Wells, Richard D. Hall, Ryan D. Gentzler, and Laura E. Barnes

Exploring the Design Space of Technological Interventions for Menopause: A Systematic Review 584
Kamala Payyapilly Thiruvenkatanathan, Shaowen Bardzell, and Jeffrey Bardzell

Personalized Healthcare

Understanding Barriers of Missing Data in Personal Informatics Systems 603
Nannan Wen, Aditi Mallavarapu, Jacob Biehl, Erin Walker, and Dmitriy Babichenko

"I Have to Do Something About It" - An Exploration of How Dashboards Invoke Self-Reflections in Chronic Obstructive Pulmonary Disease Patients 619
Stephanie Githa Nadarajah, Peder Walz Pedersen, Milo M. Skovfoged, Hamzah Ziadeh, and Hendrik Knoche

Design of a Social Chatbot with Gamification for user Profiling and Smoking Trigger Detection 636
Kyana Bosschaerts, Jeroen Stragier, Bram Steenwinckel, Lieven De Marez, Sofie Van Hoecke, and Femke Ongenae

Specification of Quality of Context Requirements for Digital Phenotyping Applications 651
Luís Eduardo Costa Laurindo, Ivan Rodrigues de Moura, Luciano Reis Coutinho, and Francisco José da Silva e Silva

Patient Data Work with Consumer Self-tracking: Exploring Affective and Temporal Dimensions in Chronic Self-care 666
Tariq Osman Andersen, Jonas Fritsch, and Stina Matthiesen

Personalizing mHealth Persuasive Interventions for Physical Activity: The Impact of Personality on the Determinants of Physical Activity 681
Alaa Alslaity, Najla Amutari, and Rita Orji

Author Index 699

Personal Informatics and Wearable Devices

Robust Respiration Sensing Based on Wi-Fi Beamforming

Wenchao Song[1], Zhu Wang[1(✉)], Zhuo Sun[1], Hualei Zhang[1], Bin Guo[1],
Zhiwen Yu[1], Chih-Chun Ho[2], and Liming Chen[3]

[1] Northwestern Polytechnical University, Xi'an, China
`wangzhu@nwpu.edu.cn`
[2] Beijing Jizhi Digital Technology Co., Ltd., Beijing, China
[3] Ulster University, Newtownabbey, UK

Abstract. Currently, the robustness of most Wi-Fi sensing systems is very limited due to that the target's reflection signal is quite weak and can be easily submerged by the ambient noise. To address this issue, we take advantage of the fact that Wi-Fi devices are commonly equipped with multiple antennas and introduce the beamforming technology to enhance the reflected signal as well as reduce the time-varying noise. We adopt the dynamic signal energy ratio for sub-carrier selection to solve the location dependency problem, based on which a robust respiration sensing system is designed and implemented. Experimental results show that when the distance between the target and the transceiver is 7 m, the mean absolute error of the respiration sensing system is less than 0.729 bpm and the corresponding accuracy reaches 94.79%, which outperforms the baseline methods.

Keywords: Beamforming · Respiration Sensing · Robustness · Wi-Fi

1 Introduction

Internet of Things (IoT) technologies are increasingly used in smart homes, smart cities, etc. Among these technologies, wireless sensing has attracted much attention from both the academic and industrial fields, due to its advantages of non-intrusive and privacy-preserving.

There are a variety of wireless signals in our daily life, among which Wi-Fi is the most common one in indoor environments. *Cisco Annual Internet Report (2018–2023)* shows that the number of public Wi-Fi hotspots will be nearly 628 millions by 2023, 4 times more than that in 2018. Therefore, wireless sensing based on the Wi-Fi signal is promising to achieve truly seamless access, attracting a great deal of research and leading to lots of breakthroughs. In 2000, Bahl et al. [1] proposed to use RSS for indoor localization and implemented Radar based on Wi-Fi for the first time. Youssef et al. [21] designed a high-precision positioning system named Horus, whose average positioning error reached 0.6 m. However, RSS is inaccurate and susceptible to multi-path effects, and thus cannot be used

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 3–17, 2023.
https://doi.org/10.1007/978-3-031-34586-9_1

for fine-grained sensing. In 2011, Halperin et al. [3] released the CSI Tool, which enables the extraction of CSI (i.e., Channel State Information) from commercial Wi-Fi devices (e.g., the Intel 5300 NIC). Compared with RSS, CSI is more fine-grained and sensitive to the environment, and has been widely used for human behavior sensing, such as trajectory tracking [4,8,18,20,23], respiration sensing [7,9–12,17,23–28], gait detection [13,15,19,22], gesture recognition [2,5–7,16].

However, most existing Wi-Fi sensing systems are still in the lab stage and there are some major issues preventing their practical deployment. One of the issues is that the sensing range of Wi-Fi based systems is very limited, not even effectively cover a room, as the target's reflection signal is quite weak. Another issue is location dependency, i.e. the sensing performance degrades greatly once the target's location varies. Thereby, how to effectively overcome the above problems and implement a robust Wi-Fi sensing system is a prominent challenge. To this end, we take advantage of the fact that Wi-Fi devices are commonly equipped with multiple antennas and introduce beamforming to enhance the target's reflection signal, which help improve the system's sensing ability in long-range scenario. Meanwhile, to address the location dependency problem, we propose an sub-carrier selection algorithm based on the dynamic signal energy ratio.

The main contributions of this paper are as follows:

- We built a beamforming-based Wi-Fi sensing platform using commercial Wi-Fi devices, based on which both delay and sum beamforming algorithms were implemented.
- To address the location dependency problem, we proposed an optimal sub-carrier selection algorithm based on the dynamic signal energy ratio.
- Based on the built Wi-Fi sensing platform, we implemented a robust respiration sensing system, which outperforms the state-of-the-art baseline methods.

The rest of the paper is organized as follows. In Sect. 2, we discuss the relevant work briefly. Then the beamforming algorithm and the respiration sensing system are presented in Sect. 3 and Sect. 4, respectively. In Sect. 5, we evaluate the performance of the proposed system. Finally, the paper is concluded in Sect. 6.

2 Related Work

This section presents recent advances in the field of respiration detection based on wireless sensing.

Sensing Range: Generally, the solutions to the limited sensing range can be divided into two categories according to the processing stage [14]. One is to strength the signal at the source. For example, LoRa, a signal used for long-range communication in the IoT, was used for sensing in [27]. The other is to enhance the received signal with proper processing. For instance, FarSense [26] assumed that the time-varying random phase shifts are similar at different antennas on the same receiver, and eliminate the noise by dividing the CSI readings.

However, the CSI obtained from either receiving antenna contains target motion information, leading to the destruction of phase information after dividing. EMA [23] proposed the sensing signal-to-noise ratio (SSNR) to measure the sensing ability of received signal, and combined the CSIs on multiple antennas by the optimal weight vector that maximize SSNR to enhance them.

Robustness: FullBreath [24] analyzed the widely existing location dependency problem using only magnitude or phase information, and proposed to use the complementary of magnitude and phase to solve the problem. Zhang et al. presented a new idea of using the curvature curve of target reflection signal to solve the location dependency problem in [27], which also alleviated the problem of insufficient spatial resolution due to the limited number of antennas. MultiSense [25] took the difference of sensing ability of sub-carriers into account, and performed a sub-carrier filtering and merging algorithm based on respiration energy to improve the robustness.

3 Beamforming

Beamforming, also known as spacial filtering, is to adjust the amplitude and phase of multiple signals so that the signals interfere with each other in the desired direction to be enhanced or weakened. There are many ways to implement beamforming, the most common of which is based on the antenna array.

3.1 Preliminary

Antenna array is a set of antennas arranged according to certain rules. The simplest of them is the equally spaced line array (hereafter referred to as line array) in which the array elements are arranged at equal intervals on a straight line, as shown in Fig. 1.

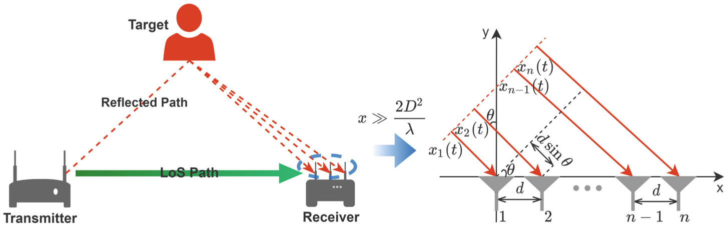


Fig. 1. Application scenario and structure of equally spaced linear array.

Let the antenna interval of the n-element line array be d, the wavelength of the signal be λ, and the angle of arrival be θ. In the far field, the received signals can be expressed as

$$
\begin{aligned}
y_1 &= \delta(t)e^{j\phi(t)}(h_{s,1} + se^{-j\frac{2\pi(1-1)d\sin\theta}{\lambda}} + \varepsilon_1) \\
y_2 &= \delta(t)e^{j\phi(t)}(h_{s,2} + se^{-j\frac{2\pi(2-1)d\sin\theta}{\lambda}} + \varepsilon_2) \\
&\vdots \\
y_n &= \delta(t)e^{j\phi(t)}(h_{s,n} + se^{-j\frac{2\pi(n-1)d\sin\theta}{\lambda}} + \varepsilon_n)
\end{aligned}
\tag{1}
$$

where $h_{s,k}$ is the static path signal, s and $e^{-j\frac{2\pi(k-1)d\sin\theta}{\lambda}}$ are the complex coefficient including amplitude and common phase, and the phase difference of the target reflected signals, respectively, ε_k is the additive Gaussian white noise (AWGN), $\delta(t)$ is the automatic gain noise (AGN), and $\phi(t)$ is the random phase offset.

To facilitate the description, we have the following equation

$$
\boldsymbol{Y} = \delta(t)e^{j\phi(t)}(\boldsymbol{H_s} + \boldsymbol{A}s + \boldsymbol{N}) \tag{2}
$$

where $\boldsymbol{Y} = [y_1, y_2, \cdots, y_n]^{\mathrm{T}}$, $\boldsymbol{H_s} = [h_{s,1}, h_{s,2}, \cdots, h_{s,n}]^{\mathrm{T}}$, $\boldsymbol{N} = [\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n]^{\mathrm{T}}$, $\boldsymbol{A} = [1, e^{-j\frac{2\pi d\sin\theta}{\lambda}}, \cdots, e^{-j\frac{2(n-1)\pi d\sin\theta}{\lambda}}]^{\mathrm{T}}$.

3.2 Delay and Sum Beamforming (DSB)

In order to enhance the signals, what we need to do is to find a proper weight vector $\boldsymbol{W} = [w_1, w_2, \cdots, w_n]^{\mathrm{T}}$ to combine the CSIs. The most comprehensible one is Delay and Sum Beamforming, which is based on the vector addition principle i.e. $\mid \boldsymbol{a} + \boldsymbol{b} \mid \leq \mid \boldsymbol{a} \mid + \mid \boldsymbol{b} \mid$, and the equal sign holds when and only when $\boldsymbol{a}$ and $\boldsymbol{b}$ are in same direction. Therefore, in order to enhance the signal, we need to perform the following steps: 1) align the signal; 2) sum them up.

Note that it is the target reflection signal that we want to enhance. As can be seen from Eq. 2, the phase differences of the target reflection signals are determined by $\boldsymbol{A}$, so the simplest way to align them is to multiply them by $\boldsymbol{A}^{\mathrm{H}}$. Then all we need to do is sum them up. In general, we can take the weight vector as

$$
\boldsymbol{W_{DSB}} = e^{-j\alpha} \cdot \boldsymbol{A}^{\mathrm{H}} = e^{-j\alpha} \cdot [1, e^{j\frac{2\pi d\sin\theta}{\lambda}}, \cdots, e^{j\frac{2(n-1)\pi d\sin\theta}{\lambda}}] \tag{3}
$$

The following is a simple example to illustrate the effect of the beamforming algorithm. As shown in Fig. 2, two original CSIs are shown on the left and the beamformed one is on the right, where green arrows are the static path signals, red ones are the target reflection signals, and purple ones are the ambient noises. The two CSIs on the left side are $y_1 = h_{s,1} + h_{d,1} + \varepsilon_1 = (6 + 3i) + (-1 + 4i) + (-1.41 - 1.41i)$ and $y_2 = h_{s,2} + h_{d,2} + \varepsilon_2 = (-4 - 8i) + (5 + i) + (2i)$, where $h_{d,2}$ and $h_{d,1}$ has a phase difference of $\delta\phi = 92.7° = 1.61rad$. According to Eq. 2, we have the weight vector $\boldsymbol{W_{DSB}} = [1, e^{j\frac{2\pi d\sin\theta}{\lambda}}] = [1, e^{j\frac{2\pi d\sin 1.61}{\lambda}}]$, which rotates the second CSI counterclockwise by 1.61 rad in the complex plane, thus aligning the dynamic vectors of CSIs. Then we sum the vectors and get $y_{comb} = h_{s,comb} + h_{d,bomb} + \varepsilon_{comb} = (14 - 0.61i) + (-2.24 + 8.95i) + (-3.41 - 1.51i)$, where $h_{d,comb}$ is 2.24 times as large as $h_{d,1}$, $h_{s,comb}$ is 2.08 times as large as $h_{s,1}$

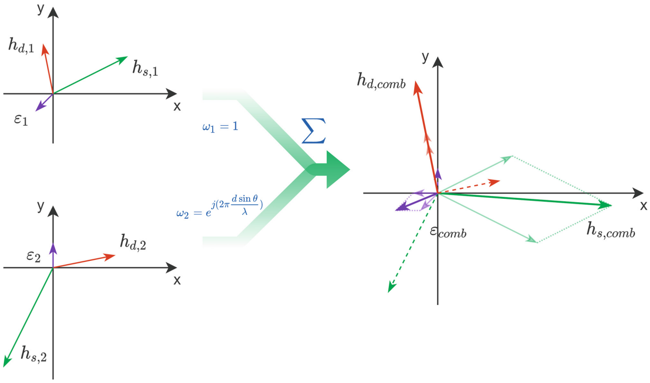


Fig. 2. A simple example of Delay and Sum Beamforming. (Color figure online)

and ε_{comb} is 1.87 times as large as ε_1, respectively. Obviously, the dynamic vector increases most significantly, which means that the dynamic path signal is enhanced and our approach works.

3.3 Beam Nulling

As shown in Eq. 2, the actual CSI also contains time-varying random phase shifts such as Carrier Frequency Offset (CFO), Sampling Frequency Offset (SFO), and automatic gain noise, components that invalidate the sensing methods based on CSI timing. FarSense [26] in which the CSI of the two antennas are divided effectively eliminate these noises, but related studies [25,27] demonstrated that the phase of the sensed signal is corrupted due to the dynamic information contained in the denominator. In the following, we present a beamforming-based method (hereinafter referred to as Beam Nulling) to eliminate the above noise without destroying the phase information.

In the previous subsection, we used a weight vector to enhance the target reflection signal by the DSB algorithm. Here, we can also use another weight vector to achieve our goal, but note that here we are tring to nulling it instead of enhancing. Therefore, we can describe the Beam Nulling problem as follows.

$$\begin{aligned} &y_{null} = \delta(t)e^{j(\phi_c+\phi_s)}(\boldsymbol{W_{null}}\boldsymbol{H_s} + \boldsymbol{W_{null}}\boldsymbol{A}s + \boldsymbol{W_{null}}\boldsymbol{N}) \\ &\boldsymbol{s.t.} \quad \boldsymbol{W_{null}}\boldsymbol{A} = 0 \end{aligned} \tag{4}$$

Analytical solution of the above problem cannot be found because the unknown dynamic path signal is only a part of the CSI and is unknown. However, we can find an approximate solution instead. To obtain an approximate solution, we introduce a metric named dynamic signal energy ratio as an indicator to solve the problem using a stochastic optimization algorithm.

Definition 1. *Dynamic signal energy ratio (DSER) is the ratio of the energy of the dynamic path signal to the total energy in the spectrum. The calculation consists of two steps:*

1. *Perform Fast Fourier Transform (FFT) within a time window to obtain the spectrum of different signal components;*
2. *Calculate the ratio of the energy of the dynamic components to the total energy.*

According to the above definition, the Beam Nulling problem can be transformed into a dynamic signal energy minimization problem. The algorithm is described as follows

Algorithm 1. Beam Nulling Algorithm

```
Input: $CSI$: CSIs of multiple antennas in one time window; $fs$: Sampling frequency;
    $f_l$: Dynamic component frequency lower bounds; $f_h$: Dynamic component frequency
    upper bounds;
Output: optimal $\boldsymbol{W_{null}}^*$;
random initial $\boldsymbol{W}^{\boldsymbol{0}}_{\boldsymbol{null}}$;
repeat
    combine signals with weights $\boldsymbol{W}^{\boldsymbol{k}}_{\boldsymbol{null}}$, i.e. $\boldsymbol{Y}^{\boldsymbol{k}} = \boldsymbol{W}^{\boldsymbol{k}}_{\boldsymbol{null}} \cdot CSI$;
    apply fast fourier transform to $\boldsymbol{Y}^{\boldsymbol{k}}$ to obtain the energies $\boldsymbol{E}^{\boldsymbol{k}}$ of each component
        and corresponding frequencies $\boldsymbol{F}$;
    calculate dynamic signal energy ratio $DSER^k = \frac{\sum_{f_l \le F_i \le f_h} E_i^k}{\sum E_i^k}$;
    calculate the value of $\boldsymbol{W}^{\boldsymbol{k+1}}_{\boldsymbol{null}}$;
until ($\left|\boldsymbol{W}^{\boldsymbol{k+1}}_{\boldsymbol{null}} - \boldsymbol{W}^{\boldsymbol{k}}_{\boldsymbol{null}}\right| < 10^{-5}$)
```

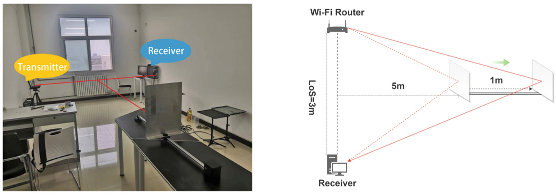
After obtaining the approximate most weight matrix $\boldsymbol{W_{null}}^*$, a reference signal without contain dynamic information and applied to eliminate random phase shift and automatic gain noise, that is

$$
\begin{aligned}
y^* = \frac{y_{beam}}{\hat{y}_{ref}} &= \frac{\cancel{\delta(t)e^{j\phi(t)}}(\boldsymbol{W}^{\mathrm{H}}\boldsymbol{H}_{\boldsymbol{s}} + \boldsymbol{W}^{\mathrm{H}}\boldsymbol{A}s + \boldsymbol{W}^{\mathrm{H}}\boldsymbol{N})}{\cancel{\delta(t)e^{j\phi(t)}}(\boldsymbol{W}^{*}_{\boldsymbol{null}}\boldsymbol{H}_{\boldsymbol{s}} + \boldsymbol{W}^{*}_{\boldsymbol{null}}\boldsymbol{N})} \\
&\approx \frac{\boldsymbol{W}^{\mathrm{H}}\boldsymbol{H}_{\boldsymbol{s}}}{\boldsymbol{W}^{*}_{\boldsymbol{null}}\boldsymbol{H}_{\boldsymbol{s}}} + \frac{\boldsymbol{W}^{\mathrm{H}}\boldsymbol{A}}{\boldsymbol{W}^{*}_{\boldsymbol{null}}\boldsymbol{H}_{\boldsymbol{s}}}s + \frac{\boldsymbol{W}^{\mathrm{H}}\boldsymbol{N}}{\boldsymbol{W}^{*}_{\boldsymbol{null}}\boldsymbol{H}_{\boldsymbol{s}}}
\end{aligned}
\tag{5}
$$

3.4 Verification of Beamforming

After the above processing, we enhance the target reflection signal and eliminate the random phase shift and auto gain noise to obtain a signal with higher sensing ability. In this subsection, we verify the effectiveness of the beamforming and beam nulling algorithms by a sliding experiment.

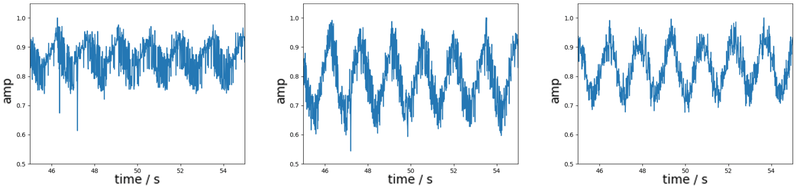
Figure 3(a) and Fig. 3(b) are the experimental scenario and setting for slide experiment, respectively. In this experiment, we put the slide with a length of 1 m on the vertical bisector of the transceiver at 5–6 m, and let the slider with the thick metal plate move slowly and uniformly from 5 m to 6 m and collect CSI packets. The change of the length of the reflection path $\varDelta x = 2 \times (\sqrt{6^2 + 1.5^2} - \sqrt{6^2 + 1.5^2}) = 1.929\,\mathrm{m}$. According to the Fresnel zone model, it is known that the metal plate theoretically crosses $\frac{\varDelta x}{\frac{\lambda}{2}} = 74.4$ boundaries of the Fresnel zone and corresponding to a waveform with 37.2 periods, where the wavelength of the 5785 MHz signal is $\lambda = \frac{3 \times 10^8}{5785 \times 10^6} = 0.0519\,\mathrm{m}$. Then we take 1000 of the 5990 CSI packets collected, which corresponding to a signal with $\frac{1000}{5990} \times 37.2 = 6.2$ periods theoretically, and process them with our proposed approach. Below we give the experimental results.



(a) (b)

Fig. 3. Experimental scenario (a) and setting (b) for slide experiment.

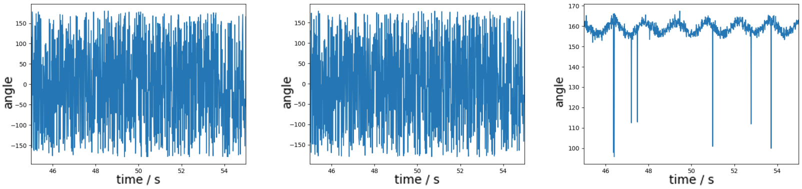
As shown in Fig. 4, it is obvious that the amplitude increases significantly in Fig. 4(b) and Fig. 4(c), which indicates that the beamforming enhances the intensity of the target reflected signal. In addition, the patterns of the signals in Fig. 4(b) and Fig. 4(c) are identical to those in Fig. 4(a) and have 6 full periods included, which is consistent with the theoretical value.



(a) (b) (c)

Fig. 4. Effect of beamforming on signal amplitude, (a), (b), (c) are the waveforms of the amplitude of the original CSI, beamformed CSI and beam-nulled CSI, respectively.

We also analyze the effect of beamforming on phase, as shown in Fig. 5. Unlike the amplitude results, the phase waveforms of both the original and beamformed signals are quite heterogeneous and do not reflect any motion pattern, but the beam-nulled signal in Fig. 5(c) shows an obvious periodicity as amplitude waveform in Fig. 4(a).



(a) (b) (c)

Fig. 5. Effect of beamforming on signal phase, (a), (b), (c) are the waveforms of the amplitude of the original CSI, beamformed CSI and beam-nulled CSI, respectively.

The above analysis leads us to the following conclusions. Our proposed beamforming algorithm can effectively enhance the target reflection signal while maintaining the pattern and periodicity of the original signal. What's more, the beam nulling algorithm in this paper can recover the phase information of the signal from the noisy original signal.

4 Respiration Sensing System

This section first introduces the mechanism and solution of position dependency, and then introduces the respiration sensing system proposed in this paper.

4.1 Location Dependency

Currently, most of the respiration sensing algorithms are based on the amplitude. Considering the original signal phase information in the previous section, this phenomenon is not difficult to understand. However, this approach suffers from a serious location dependency problem, i.e., the sensing ability of the signal declines severely when the target is only at different locations which causes the same signal change.

Figure 6 uncovers the reason for position dependency problem. Although both the static path signal (green arrow) and target reflected signal (red or yellow arrow) changes are the same in the above figure, but the amplitude of the fluctuations of the synthesized signal (blue arrow) varies greatly. The amplitude fluctuation of the synthesized signal is the smallest when the change range of the target reflection signal is symmetric about the static path signal while the amplitude fluctuation of the synthesized signal is the largest when the change

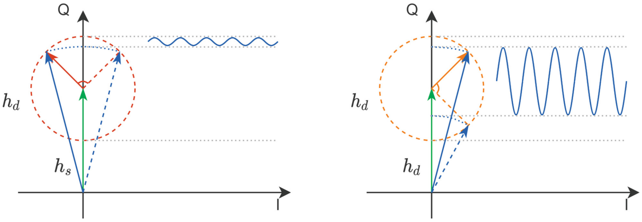


Fig. 6. The mechanism underlying the location dependence problem. (Color figure online)

range of the target reflection signal is symmetric about the vertical line of the static path signal.

The fundamental reason is that the synthetic signal used for sensing is influenced by the static path signal in addition to the target reflected one. Therefore, a possible way to solve this problem is to remove the static path signal from the synthesized signal. To remove the static component, we do band-pass filtering in frequency domain, retaining only the frequency components in the respiration band, thus solving the location dependency problem by using the location-independent target reflected signal features.

4.2 Respiration Sensing

We depict the respiration sensing process in Fig. 7, which is divided into three main steps, i.e. pre-processing, respiration feature extraction and target detection.

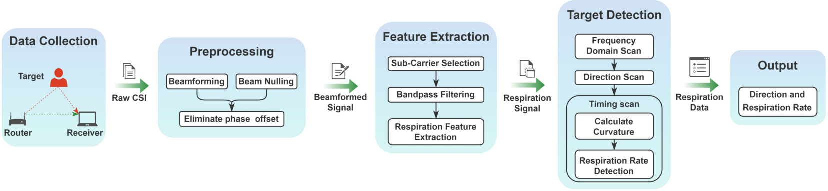


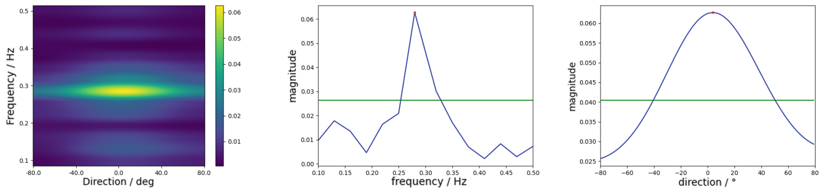
Fig. 7. The process of respiration sensing algorithm.

Pre-processing The purpose of pre-processing is to enhance the target reflection signal, reduce noise and select the optimal sub-carrier.

For each sub-carrier of the original CSI, a angle range from $-80°$ to $80°$ is scanned in steps of $1°$. At each scanning step, a weight vector of DSB algorithm is computed, and then the CSIs are weighted and summed to obtain the synthetic signal. After that, we take 30 s as the time window, 0.1 ˜0.5 Hz as the respiration belt, using Algorithm 1 to approximate the optimal weights and constructing a reference vector to eliminate time-varying errors.

Breathing Feature Extraction. We found that the sensing ability of different sub-carriers are quite different through a large number of experiments. Thus *DSER* is used as an indicator to select the sub-carrier with the strongest respiration energy for the following steps to ensure the quality of respiration features.

In order to extract the respiration features, we perform band-filtering on the optimal sub-carrier, and only retain the energy of the components in the respiration frequency band, so as to obtain the frequency-direction spectrum as in Fig. 8(a) and the filtered signals. As Fig. 8(a) shown, there is a target in the 5° direction with a respiration frequency of 0.3 Hz.



(a) frequency-direction (b) energy-frequency (c) energy-direction

Fig. 8. An spectrum example of respiration sensing. (Color figure online)

Target Detection. To avoid the effect of side lobes, our target detection algorithm is divided into three steps, namely frequency scanning, direction scanning and timing detection, which means that we simultaneously utilize the information of three dimensions of frequency, space and time.

When scanning the frequency, we extract the maximum energy of a certain frequency in all directions within the respiration range as the energy of the frequency. Then a reasonable threshold that is calculated with the maximum and minimum energies is set to filter out the energy peaks greater than it to obtain candidate frequencies. For example, we get a candidate frequency of 0.3Hz in Fig. 8(b), where the green line is the threshold.

For each candidate frequency, the next step is to scan each direction. We extract the energy-direction curve from the direction-frequency spectrum and perform the same way as frequency scanning to obtain candidate directions. We get a candidate direction of 5° in Fig. 8(c).

For each candidate direction of each candidate frequency, we take out the filtered signal and use the short-term auto-correlation function to obtain the corresponding respiration rate.

After performing the above steps, our system will output the target's respiration rate.

5 Evaluation

5.1 Experiment Setup

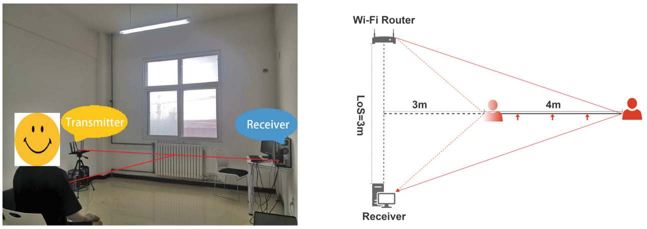
Our system uses a TP-Link wireless router as the transmitter, whose frequency is set to 5785 MHz, the receiver is a PC equipped with an Intel 5300 NIC and three omnidirectional antennas, and CSI Tool in [3] is used to collect CSI. During the experiment, both the transmitter and receiver are at a height of 85cm, and subject breathes naturally with a metronome.

The experimental parameters were set as follows unless otherwise specified. The sampling frequency is set 100 Hz, the distance between transceiver is 3 m, and the subject is located at 4 m on the vertical bisector of the transceiver. We collect 150 s of CSI at each sampling point and process it in segments.

In order to evaluate the system, the indicators used are described here. Target distance is half of the sum of the distance between the target and the transceiver, i.e. $x = \sqrt{d^2 + \frac{LoS}{2}^2}$. The mean absolute error (MAE) is the average of the absolute errors of the predicted respiratory rate for all segments, and accuracy is $(1 - \frac{MAE}{BPM_{truth}}) \times 100\%$.

5.2 Comprehensive Experiment

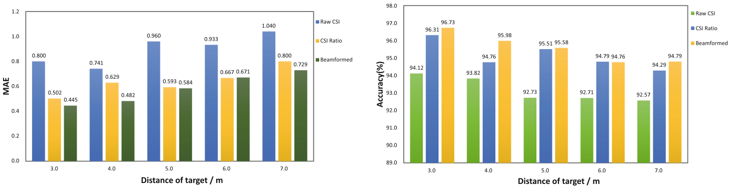
As Fig. 9 shown, in this set of experiments, let the target located at a distance d, where d is taken as 3 m, 4 m, 5 m, 6 m, and 7 m. To evaluate our method, we use the raw CSI, CSI quotient and beamformed CSI as inputs, respectively, and calculate the subject's respiration rate with the same period detection algorithm. Finally, we have the following results.



(a) (b)

Fig. 9. Experimental scenario and experimental setup for comprehensive experiment.

From the Fig. 10, it can be seen that the MAE of both beamformed CSI and CSI ratio is much lower (or the accuracy is much higher than that of original CSI), indicating that our method is as effective as CSI ratio in enhancing signal sending ability. The MAE of beamforming is 0.445 BPM (breath per minute) when the target is located at 3 m on the vertical bisector of the transceiver, Though, when the distance increases to 7 m, the MAE of the original CSI exceeds 1 BPM while that of beamforming in this paper is still far less than 1 BPM ans is 0.729 BPM, which well meet the accuracy requirements in application scenarios.

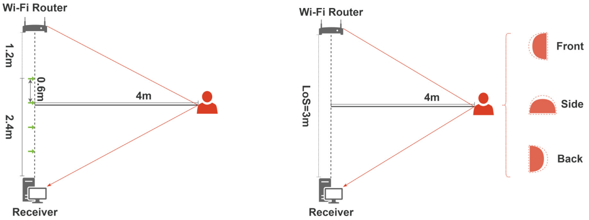


(a) MAE (b) Accuracy

Fig. 10. Experimental results of comprehensive experiment

5.3 Other Experiments

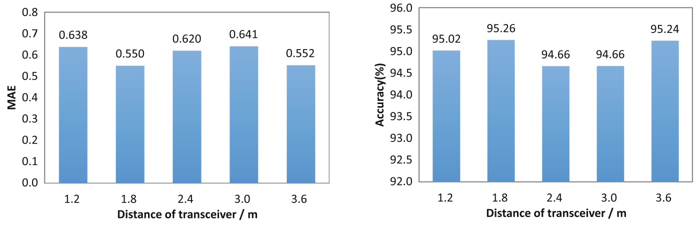
We also set up experiments to investigate the effect of transceiver distance and target orientation on sensing performance, and the experimental setup is shown in Fig. 11(a) and Fig. 11(b).



(a) Varying the distance (b) Varying the orientation

Fig. 11. Experiment setup of different transceiver distances and target orientations

Effect of Transceiver Distance. In this set of experiments, we explore the impact of LoS path signals on system performance by changing the transceiver distance from 1.2 m to 3.6 m in steps of 0.6 m, as shown in Fig. 11(a).

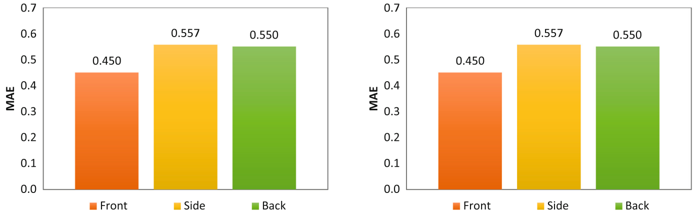


(a) MAE (b) Accuracy

Fig. 12. Experimental results with different transceiver distances

As shown in Fig. 12, it can be seen that both MAE and accuracy tend to be a random distribution, which means that the LoS signal strength does not affect the performance of the sensing system. Such experimental results is consistent with the analysis in [23].

Effect of Target Orientation. As shown in Fig. 13, the system has the smallest average absolute error and the highest accuracy when the front of the body faces the transceiver, while the MAE is relatively large (or the accuracy is slightly lower) when the side and back face the transceiver. This result is expected because the displacement of the front side of a human chest is the largest when breathing, which leads to a more obvious variation on the signal, making the accuracy higher; in contrast, the displacement of the back and side of the chest is smaller, making the accuracy lower.



(a) MAE (b) Accuracy

Fig. 13. Experimental results of target orientation

6 Conclusion

In this paper, to achieve robust respiration detection, we introduced the beamforming technology as well as a metric named dynamic signal energy ratio into the field of Wi-Fi sensing. Specifically, the commonly used delay and sum beamforming algorithm is implemented and its practical effects is verified, based on which a Wi-Fi beamforming-based sensing system is developed. The performance of the proposed system is evaluated with a number of experiments, and results show that the system achieves high accuracy and robustness, compared with baseline methods.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China (No. 61960206008, 62072375, 62102322), and the Fundamental Research Funds for the Central Universities (No. D5000210786).

References

1. Bahl, P., Padmanabhan, V.N.: Radar: an in-building RF-based user location and tracking system. In: Proceedings of IEEE INFOCOM, vol. 2, pp. 775–784. IEEE (2000)
2. Gao, R., et al.: Towards robust gesture recognition by characterizing the sensing quality of WiFi signals. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **6**(1), 1–26 (2022)
3. Halperin, D., Hu, W., Sheth, A., Wetherall, D.: Tool release: gathering 802.11 n traces with channel state information. ACM SIGCOMM Comput. Commun. Rev. **41**(1), 53 (2011)
4. Li, X., et al.: IndoTrack: device-free indoor human tracking with commodity Wi-Fi. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(3), 1–22 (2017)
5. Liu, J., Wang, Y., Chen, Y., Yang, J., Chen, X., Cheng, J.: Tracking vital signs during sleep leveraging off-the-shelf WiFi. In: Proceedings of ACM MobiHoc, pp. 267–276 (2015)
6. Liu, X., Cao, J., Tang, S., Wen, J.: Wi-sleep: contactless sleep monitoring via WiFi signals. In: 2014 IEEE Real-Time Systems Symposium, pp. 346–355. IEEE (2014)
7. Niu, K., Zhang, F., Xiong, J., Li, X., Yi, E., Zhang, D.: Boosting fine-grained activity sensing by embracing wireless multipath effects. In: Proceedings of ACM CoNEXT, pp. 139–151 (2018)
8. Qian, K., Wu, C., Yang, Z., Liu, Y., Jamieson, K.: Widar: decimeter-level passive tracking via velocity monitoring with commodity Wi-Fi. In: Proceedings of ACM MobiHoc, pp. 1–10 (2017)
9. Wang, H., et al.: Human respiration detection with commodity WiFi devices: do user location and body orientation matter? In: Proceedings of ACM UbiComp, pp. 25–36 (2016)
10. Wang, P., Guo, B., Xin, T., Wang, Z., Yu, Z.: TinySense: multi-user respiration detection using Wi-Fi CSI signals. In: IEEE 19th International Conference on e-Health Networking, Applications and Services, pp. 1–6 (2017)
11. Wang, X., Yang, C., Mao, S.: PhaseBeat: exploiting CSI phase data for vital sign monitoring with commodity WiFi devices. In: IEEE 37th International Conference on Distributed Computing Systems (ICDCS), pp. 1230–1239. IEEE (2017)
12. Wang, X., Yang, C., Mao, S.: TensorBeat: tensor decomposition for monitoring multiperson breathing beats with commodity WiFi. ACM Trans. Intell. Syst. Technol. (TIST) **9**(1), 1–27 (2017)
13. Wang, Z., Guo, B., Yu, Z., Zhou, X.: Wi-Fi CSI-based behavior recognition: from signals and actions to activities. IEEE Commun. Mag. **56**(5), 109–115 (2018)

14. Wang, Z., Yu, Z., Lou, X., Guo, B., Chen, L.: Gesture-radar: a dual doppler radar based system for robust recognition and quantitative profiling of human gestures. IEEE Trans. Hum.-Mach. Syst. **51**(1), 32–43 (2021)
15. Wu, C., Zhang, F., Hu, Y., Liu, K.R.: GaitWay: monitoring and recognizing gait speed through the walls. IEEE Trans. Mob. Comput. **20**(6), 2186–2199 (2020)
16. Wu, D., et al.: FingerDraw: sub-wavelength level finger motion tracking with WiFi signals. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **4**(1), 1–27 (2020)
17. Wu, D., Zhang, D., Xu, C., Wang, H., Li, X.: Device-free WiFi human sensing: from pattern-based to model-based approaches. IEEE Commun. Mag. **55**(10), 91–97 (2017)
18. Wu, D., Zhang, D., Xu, C., Wang, Y., Wang, H.: WiDir: walking direction estimation using wireless signals. In: Proceedings of ACM UbiComp, pp. 351–362 (2016)
19. Xin, T., Guo, B., Wang, Z., Li, M., Yu, Z., Zhou, X.: FreeSense: indoor human identification with Wi-Fi signals. In: Proceedings of IEEE GLOBECOM, pp. 1–7 (2016)
20. Xin, T., et al.: FreeSense: a robust approach for indoor human detection using Wi-Fi signals. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **2**(3) (2018)
21. Youssef, M., Agrawala, A.: The Horus WLAN location determination system. In: Proceedings of ACM MobiSys, pp. 205–218 (2005)
22. Yu, N., Wang, W., Liu, A.X., Kong, L.: QGesture: quantifying gesture distance and direction with WiFi signals. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **2**(1), 1–23 (2018)
23. Zeng, Y., Liu, J., Xiong, J., Liu, Z., Wu, D., Zhang, D.: Exploring multiple antennas for long-range WiFi sensing. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **5**(4), 1–30 (2021)
24. Zeng, Y., Wu, D., Gao, R., Gu, T., Zhang, D.: FullBreathe: full human respiration detection exploiting complementarity of CSI phase and amplitude of WiFi signals. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **2**(3), 1–19 (2018)
25. Zeng, Y., Wu, D., Xiong, J., Liu, J., Liu, Z., Zhang, D.: MultiSense: enabling multi-person respiration sensing with commodity WiFi. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **4**(3), 1–29 (2020)
26. Zeng, Y., Wu, D., Xiong, J., Yi, E., Gao, R., Zhang, D.: FarSense: pushing the range limit of WiFi-based respiration sensing with CSI ratio of two antennas. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **3**(3), 1–26 (2019)
27. Zhang, F., et al.: Unlocking the beamforming potential of LoRa for long-range multi-target respiration sensing. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **5**(2), 1–25 (2021)
28. Zhang, H., et al.: Understanding the mechanism of through-wall wireless sensing: a model-based perspective. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **6**(4), 1–28 (2022)

Heart Rate During Sleep Measured Using Finger-, Wrist- and Chest-Worn Devices: A Comparison Study

Nouran Abdalazim[1](✉), Joseba Aitzol Arbilla Larraza[1],
Leonardo Alchieri[1], Lidia Alecci[1], Silvia Santini[1], and Shkurta Gashi[2]

[1] Università della Svizzera italiana (USI), Lugano, Switzerland
{`nouran.abdalazim,joseba.aitzol.arbilla.larraza,leonardo.alchieri,`
`lidia.alecci,silvia.santini`}`@usi.ch`
[2] ETH AI Center, Zürich, Switzerland
`shkurta.gashi@ai.ethz.ch`

Abstract. Wearable heart rate (HR) sensing devices are increasingly used to monitor human health. The availability and the quality of the HR measurements may however be affected by the body location at which the device is worn. The goal of this paper is to compare HR data collected from different devices and body locations and to investigate their interchangeability at different stages of the data analysis pipeline. To this goal, we conduct a data collection campaign and collect HR data from three devices worn at different body positions (finger, wrist, chest): The Oura ring, the Empatica E4 wristband and the Polar chestbelt. We recruit five participants for 30 nights and gather HR data along with self-reports about sleep behavior. We compare the raw data, the features extracted from this data over different window sizes, and the performance of models that use these features in recognizing sleep quality. Raw HR data from the three devices show a high positive correlation. When features are extracted from the raw data, though, both small and significant differences can be observed. Ultimately, the accuracy of a sleep quality recognition classifier does not show significant differences when the input data is derived from the Oura ring or the E4 wristband. Taken together, our results indicate that the HR measurements collected from the considered devices and body locations are interchangeable. These findings open up new opportunities for sleep monitoring systems to leverage multiple devices for continuous sleep tracking.

Keywords: Heart Rate · Wearable Devices · Ring · Wristband · Chestbelt · Statistical Analysis · Sleep Monitoring · Sleep Quality Recognition

1 Introduction

Personal health monitoring systems have recently received significant attention. They are capable of providing continuous and real time feedback to users about

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 18–32, 2023.
https://doi.org/10.1007/978-3-031-34586-9_2

their health and daily behaviour [42]. Such systems rely on different physiological signals, such as, e.g., heart rate, to assess users' health state.

Improvements in sensors, battery and storage of wearable devices make them more powerful, affordable and pervasive. These improvements increase their ability to capture various physiological signals which help in return the development of personal health monitoring systems [44]. Such evolution encourages their employment in many health domains. Most wearables are capable of capturing heart rate (HR) traces, which can be employed in many health related applications like monitoring human stress [30,43], recovery after exercise [17] and sleep behaviour [32,44]. Changes in HR and heart rate variability (HRV) reflect autonomic nervous system patterns [44] and have been correlated with sleep stages [45,46], stress [31,52] and affect [47].

The availability of several health monitoring devices makes finding the most convenient device very challenging both for researchers and end users. The rapid development of wearables created a gap between the available devices and their evaluation studies [15,50]. Therefore, a comparison is needed to determine whether the sensor readings are interchangeable between devices, placed on different body positions. While there exist a few studies that investigated the measurements of wearable devices [37,50], it is not clear whether the raw physiological data are exchangeable and how such sensor measurements perform in downstream tasks. This understanding would allow researchers to make informed decisions regarding the use of such devices in data collection studies and users to choose the device that matches their needs without hampering the quality of the measurements.

In this paper, we investigate the interchangeability of HR signals obtained from three body positions, namely, finger, wrist and chest during sleep, since one of the wearables used (Oura ring) is dedicated and provides data continuously only during dormancy. To this goal, we run a data collection campaign in the wild, to gather physiological HR data – along with self reports about sleep behavior – using three well known devices: Oura ring (generation 3), Empatica E4 wristband and Polar chestbelt. We make the dataset available to other researchers upon request and signature of a data sharing agreement. Then, we assess the interchangeability of HR collected from wearables worn at different body locations. We extensively analyze the collected data using statistical measures as well as a sleep quality recognition task, to explore the interchangeability of HR measures at the level of raw data, time-domain features and classification capability. The main contributions of this paper are as follows:

- We collect and provide to the research community a dataset[1], named **HeartS**[2] collected from five participants over 30 days in their natural environments. The dataset contains heart rate data collected using three wearable devices – Oura ring (third generation), Empatica E4 wristband and Polar chestbelt – and self-reports regarding sleep and wake up times as well as sleep quality.

---

[1] Please contact the corresponding author of the paper to make a request regarding the dataset.

[2] Heart Rate from multiple devices and body positions for Sleep measurement.

- We perform extensive statistical analysis on the raw HR data and show a high correlation between the data of the three devices, suggesting their interchangeability.
- We extract common HR features from the three devices and show that the majority of the features have statistically negligible differences in small window sizes and small differences in large window size, which may impact machine learning tasks that use such features.
- We develop a machine learning pipeline and investigate the effect of the extracted features in sleep quality recognition. Our results confirm the interchangeability of the considered devices for this task.

This paper is structured as follows. Section 2 presents an overview of the similar studies in the literature; Sect. 3 describes the data collection procedure, while Sect. 4 shows the comparison between HR signals from wearable devices. Follows Sect. 5, which presents the analysis of HR features and the machine learning task adopted for the comparison between wearable devices, and Sect. 6, which describes the limitations and future work that can be addressed. Finally, Sect. 7 presents the conclusion.

2 Related Work

Several researchers evaluate the performance of wearable devices by analyzing the provided sleep parameters, e.g., Total Sleep Time (TST), Total Wake Time (TWT), Sleep Efficiency (SE), Wake After Sleep Onset (WASO) [37,50]. Such studies are either conducted in controlled settings [44,48], or in unrestricted ones [15,37,50].

Roberts et al. [44], for instance, conduct a comparison study between consumer wearables (Oura ring, Apple watch)[3], two actigraphy wristbands and Polysomnography (PSG), used as ground truth. They report that data from commercial multi-sensor wearables are highly correlated and can be adopted in sleep-wake classification problem, competing with research-grade devices. Scott et al. [48] perform a comparison study between a new commercial smart ring (THIM) (See Footnote 3), two popular wearables (Fitbit and Actiwatch) (See Footnote 3) versus PSG. Their results show no significant differences between PSG and THIM. Other researchers evaluate the sleep-wake recognition capabilities, but do not compare neither the features nor the raw physiological signals [44,48].

Stone et al. [50] compare nine sleep tracking consumer devices positioned on wrist, finger or mattress-affixed monitors, using as ground truth Electroencephalography (ECG). Using sleep parameters, they show that Fitbit Ionic and Oura ring have the highest accuracy and minimum bias in calculating the TST,

---

[3] **Oura Ring**: https://ouraring.com; **Apple watch**: https://www.apple.com/watch/; **THIM ring:** https://thim.io; **Fitbit:** https://www.fitbit.com/; **Actiwatch:** https://www.usa.philips.com/healthcare/sites/actigraphy; **Samsung Gear Sport watch:** https://www.samsung.com/us/watches/galaxy-watch4/.

TWT and SE; while with sleep staging metrics they find no accurate result from commercial devices. Mehrabadi et al. [37] compare sleep parameters (e.g., TST, SE and WASO) from the Oura ring and the Samsung Gear Sport watch (See Footnote 3) versus a medically approved actigraphy device. They found significant correlation of both devices with actigraphy. However, neither of [37,50] applied comparisons over physiological data.

Table 1 provides a overview of the recently conducted studies that compare different wearable devices, where we can observe that only two are publicly available. The previously mentioned studies focus on specific sleep parameters, showing their interchangeability between different devices [37,48]. However, there is a gap with regards to the analysis of physiological signals, collected from various devices, and how these differences can impact sleep quality recognition task.

Table 1. Description of existing studies that compare different wearable devices

Study	Study Settings	Number of Participants	Study Duration	Publicly Available
[15]	Home	21	7 nights	No
[38]	Laboratory	6	9 nights	Yes
[48]	Laboratory	25	1 night	No
[44]	Laboratory	8	4 nights	No
[37]	Home	45	7 nights	Yes
[50]	Home	5	98 nights	No

3 Data Collection Campaign

We conduct a data collection campaign using two commercial devices and one research-grade device, for the HeartS dataset. In this section, we describe the study participants, the adopted devices, the collected data and the data collection procedure. The study is reviewed and approved by our Faculty's delegate for Ethics.

3.1 Participants

We recruit five participants (three females and two males) of age from 24 to 29 years (avg: 26.2, std: 2.3). Participants wear, for 30 consecutive nights, three wearable devices: (1) The Oura ring (Generation 3) (See Footnote 3), which measures sleep with a Photoplethysmography (PPG) sensor, from which HR and HRV are extracted [3,14]; (2) The Empatica E4 wristband[4], which is a research-grade wristband that extracts HR via PPG sensor [49]; (3) The Polar chestbelt, equipped with an Electrocardiogram (ECG) sensor [28]. Both the Polar H07[5] and

---

[4] https://www.empatica.com/en-gb/research/e4/.

[5] https://support.polar.com/e_manuals/H7_Heart_Rate_Sensor/Polar_H7_Heart_Rate_Sensor_accessory_manual_English.pdf.

H10[6] releases are included in the study, since we have only two Polar chestbelt. E4 wristband and Polar chestbelt, along with previous generations of Oura ring, are adopted in several studies in the literature, e.g., in [3,5,6,11,12,14,26,27,50]. The devices contain also other sensors, for instance, the E4 is equipped with electrodermal activity, accelerometer and skin temperature sensors. In this study, we use only the HR measurements because it is the only common sensor in all the devices, which allows us to compare them.

3.2 Data Collection Procedure

To design the study and collect the data, we follow similar procedures to the literature (e.g., [26,45,50]). At the beginning of the data collection procedure, all participants sign an informed consent form. We provide the devices, pen-and-paper diaries, and the instructions needed to set up the designated synchronization applications for obtaining the raw data from devices. We instruct participants to wear the devices on their left hand, since small lateral differences might be present if choosing difference sides [1]. Every night all participants wear the Oura ring and the E4 wristband, whereas only two wear the Polar chestbelt (since we have only two Polar chestbelts available). The participants wear the devices one hour before sleep and log the bed-time. The next day, the participants complete the self report about the sleep quality of the previous night and the wake up time then take off the devices one hour after waking up. During the day, the participants synchronize the collected data during the previous night from each device and charge the devices. To make sure of the quality and quantity of the collected data, we systematically monitor the compliance with the data collection.

3.3 Collected Data

We collect two types of data, *physiological data* using three wearable devices and *self-reports* using pen-and-paper diaries described as follows.

Physiological Data. The Oura ring provides one HR data point every five minutes during sleep as well as the *bed-time start* and the *bed-time end*. Participants use the Oura mobile application to synchronize the collected data to the Oura cloud dashboard. The Empatica E4 wristband provides HR values every second. Participants use the E4 manager desktop application[7] to synchronize the collected data to the pre-created study on the E4 website. The Polar chestbelt integrates with a third party mobile application named Polar Sensor Logger[8] to provide an HR value per second. The application stores the collected data on the device.

---

[6] https://www.polar.com/en/sensors/h10-heart-rate-sensor.

[7] https://support.empatica.com/hc/en-us/articles/206373545-Download-and-install-the-E4-manager-on-your-Windows-computer.

[8] https://play.google.com/store/apps/details?id=com.j_ware.polarsensorlogger&hl=en\&gl=US.

We collect data of 105 sleep sessions. One participant did not wear the E4 wristband on the left hand so we discard the corresponding sessions to be consistent among the participants. In total we have 98 sessions. We collect 9,038 HR data points from the Oura ring which are equivalent to about 753 h of data. For the E4 wristband, we collect 3,192,319 points (about 886 h), with mean (± standard deviation) 56.09 ± 14.58 bpm and 61.81 ± 12.42 bpm respectively. From the Polar chestbelt, we collect 656,038 HR data points, equivalent to approximately 182 h with mean 59.26 ± 12.23 bpm.

Self Reports. Participants use the pen-and-paper diaries to provide daily self reports about: their bed and wake up time, latency (i.e., the estimated time until the participant fall asleep), number of awakenings and sleep quality level every night, similar to [26,45]. They report sleep quality on a five level Likert scale [33]: *very poor, poor, normal, good, excellent* following [10]. One of the participants stopped logging self reports after the first week of the study. The dataset thus contains 80 sleep sessions labelled with the sleep behaviour.

4 Comparison of 5-Minutes Averaged HR Signals

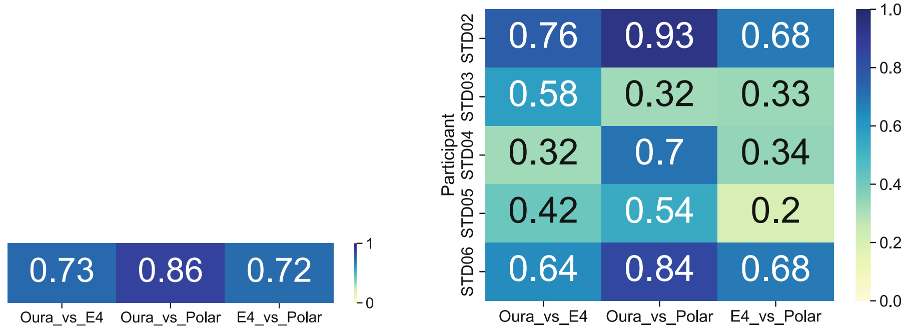
In this section we report the analysis performed using the HR signals collected as described in Subsect. 3.3. In particular, we describe the pre-processing steps, correlation and bias analysis.

4.1 Data Pre-processing

Since the Oura ring provides an average HR value every five minutes, for the current signal analysis, we down-sample the HR measurements to the same sampling frequency to obtain the same data granularity. In particular, we average the HR data of the E4 and Polar devices over five-minutes window. Given that Oura only provides the HR data during sleep, we use the *bed-time start* and *end* provided by Oura to define the sleep period and to segment the data of the E4 wristband and Polar chestbelt. We refer to the obtained traces as **averaged HR**.

4.2 Correlation Analysis

We use Shapiro-Wilk normality test to evaluate the parametric characteristic of the averaged HR [21]. We observe that the HR data, from all devices, is not normally distributed (p-value < 0.05). Based on that, we use Spearman's ρ rank correlation coefficient [35] to quantify the association between the HR signals. We conduct the analysis in two steps: first, we compute the correlation between each pair of devices using the averaged HR from all participants stacked together; then we compute the correlation using averaged HR, for each pair of device, per participant. Figure 1a shows the obtained correlation coefficients between averaged HR from every pair of devices. We find a high positive correlation



(a) Per devices Spearman's ρ correlation

(b) Per participant Spearman's ρ correlation

Fig. 1. Spearman's ρ correlation results between raw HR

between the averaged HR data from the three devices (>0.7), with the highest correlation between the Oura ring and the Polar chestbelt (0.86). All reported results are statistically significant, tested using an initial threshold of $\alpha = 0.05$ and Bonferroni [4] corrected to $\alpha_n = 0.01$, $n = 3$, as suggested in [25]. We also observe in Fig. 1b that correlation results by participant and device are similar to those by device (Fig. 1a), since they are all positive (>0.2). From this experiment, we conclude that there is a high positive correlation between averaged HR data from the three devices and the results are statistically significant ($\alpha = 0.05$, $\alpha_n = 0.003$, $n = 15$). The correlation analysis suggests interchangeability of the HR data across the three devices.

4.3 Bias Analysis

To assess the average difference between the devices, i.e., *bias* [15], we use a modified version of the Bland Altman Plot [7]. This plot measures the absolute difference between two distributions against the pair-wise averages. From the plot in Fig. 2, we observe that the data from Oura ring and the Polar chestbelt have the least average absolute difference (2.17), while the data from E4 wristband and the Polar chestbelt have the highest (5.01). These results confirm the higher correlation found between Oura's and Polar's HR data, as shown in Subsect. 4.2. In general, these results confirm the findings of the correlation analysis presented above.

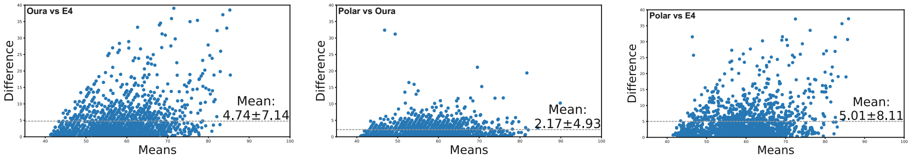


Fig. 2. Bias analysis results between every pair of devices. We report the average absolute differences and standard deviations.

5 Comparison of Features Extracted from Raw HR Signals

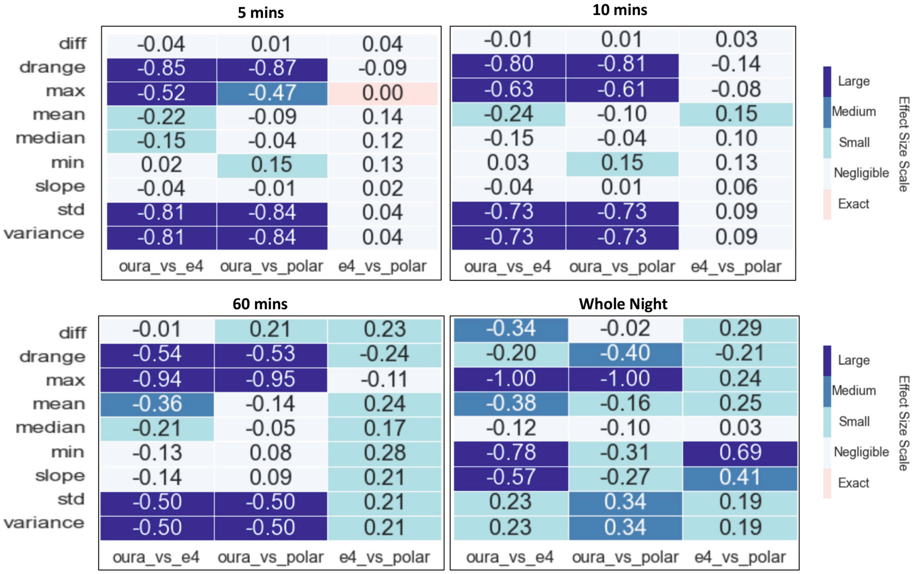
Many human activity recognition tasks rely on features extracted from the HR signals. This is in particular the case for sleep monitoring applications [32]. Thus, we extract time-domain HR features from each device using different window sizes and compare them. We execute the analysis in two steps. First, we analyze statistical differences in the extracted features. Then, we compare the performance obtained by machine learning (ML) classifiers for a sleep quality recognition task that use these features as input.

5.1 Data Pre-processing and Cleaning

In this part of data analysis, we rely on raw HR data collected from Oura ring and E4 wristband only. This is because we obtained only 18 sleep sessions for the Polar chestbelt, as opposed to approximately 80 for the other devices. We define *sleep sessions*, based on the *bed-time start* and *end* provided by Oura. We extract time-domain HR features, specifically *mean, standard deviation, range, median, variance, minimum, maximum, difference, slope*, over different window sizes, similarly to [26,39,46]. We employ three non-overlapping window sizes of 5, 10 and 60 min, and a window corresponding to the whole sleep session, similar to [25,39].

5.2 Effect Size Quantification of HR Features

To assess the difference between features extracted over each window from these devices, we employ Cliff's δ effect size [16], which allows us to determine the degree of difference between two samples. Cliff's δ values range between $[-1, 1]$, where 0 means that the two distributions are not different, while -1 and 1 indicate no distribution overlap [36]. We show the results in Fig. 3. We observe that for small window sizes, i.e., 5 and 10 min, most of the features show negligible or small differences. Also, it is noticeable that large differences are present with some Oura features due to its limited sampling rate in the designated windows. Such features rely on the data variability, e.g., the standard deviation is always 0 in a 5 min window for all Oura's data. By increasing the window size, we can



5 mins

	oura_vs_e4	oura_vs_polar	e4_vs_polar
diff	-0.04	0.01	0.04
drange	-0.85	-0.87	-0.09
max	-0.52	-0.47	0.00
mean	-0.22	-0.09	0.14
median	-0.15	-0.04	0.12
min	0.02	0.15	0.13
slope	-0.04	-0.01	0.02
std	-0.81	-0.84	0.04
variance	-0.81	-0.84	0.04

10 mins

	oura_vs_e4	oura_vs_polar	e4_vs_polar
diff	-0.01	0.01	0.03
drange	-0.80	-0.81	-0.14
max	-0.63	-0.61	-0.08
mean	-0.24	-0.10	0.15
median	-0.15	-0.04	0.10
min	0.03	0.15	0.13
slope	-0.04	0.01	0.06
std	-0.73	-0.73	0.09
variance	-0.73	-0.73	0.09

60 mins

	oura_vs_e4	oura_vs_polar	e4_vs_polar
diff	-0.01	0.21	0.23
drange	-0.54	-0.53	-0.24
max	-0.94	-0.95	-0.11
mean	-0.36	-0.14	0.24
median	-0.21	-0.05	0.17
min	-0.13	0.08	0.28
slope	-0.14	0.09	0.21
std	-0.50	-0.50	0.21
variance	-0.50	-0.50	0.21

Whole Night

	oura_vs_e4	oura_vs_polar	e4_vs_polar
diff	-0.34	-0.02	0.29
drange	-0.20	-0.40	-0.21
max	-1.00	-1.00	0.24
mean	-0.38	-0.16	0.25
median	-0.12	-0.10	0.03
min	-0.78	-0.31	0.69
slope	-0.57	-0.27	0.41
std	0.23	0.34	0.19
variance	0.23	0.34	0.19

Fig. 3. Cliff's δ effect size or HR features over different window sizes

observe that the majority of the features have small differences and the differences in the features that require data variability decrease. Such observations motivate the next experiment, where we evaluate the impact of the features' differences on a sleep recognition task. These results, similarly to the correlation experiment results in Sect. 4, suggest limited differences and interchangeability between the devices.

5.3 Sleep Quality Recognition Task

In this part of the analysis, we detail the comparison between devices in a machine learning task, with the aim of assessing the impact of the observed differences in the HR features from the devices. We employ a sleep quality recognition task, given that Oura ring is indeed dedicated to sleep behavior monitoring. As ground truth, we used the subjective sleep quality scores from the self reports, described in Subsect. 3.3.

Classification Procedure. We define this problem as a binary classification task, using normalised sleep quality labels (range [0,1]), with a threshold of 0.5 to discriminate between positive (high) and negative (low) class, similar to [26,54]. From this we obtain the following distribution: 61% of the data on the positive class and 39% on the negative class. Since the data is not completely balanced, we employ, only during training, synthetic minority over-sampling technique (SMOTE) [13]. Since the collected data provides one score per sleep session, when using 5, 10 and 60 min windows we train by assigning the label to each window. However, at validation time, we evaluate only one score per sleep sessions: to obtain this, we apply majority voting over the window-level predictions. When training using the whole sleep session, majority voting is not used.

Classification Models. We adopt 10 classification models in our experiments: Decision Tree (DT) [51], Gaussian Naïve Bayes (NB) [20], Support Vector Machine (SVM) [19], Multilayer Perceptron (MLP) [24], k-nearest neighbour (KNN) [41], Random Forest (RF) [8], XGBoost [23], AdaBoost [22], Quadratic Discriminant Analysis (QDA) [29] and Gaussian Process (GP) [53]. We use the implementation of such algorithms from the Scikit-Learn Python library [40].

Evaluation Methodology. We perform evaluation of the chosen models using two *cross validation* paradigms. We first evaluate the models' capability to generalize to a new participant, this is achieved with *Leave One Participant Out* cross validation, in which we train each model using all but one user's data, and use the remaining user's data as test set. Second, we evaluate the ability of the models to recognize the sleep quality score for an already existing participant using the *Leave One Session Out* cross validation. This procedure allows to train on all sleep sessions but one, and test on the remaining one. We use two baseline classifiers, to identify if our models are capable of learning patterns from the input data [34]. The first baseline is the Random Guess classifier (RG), which makes sleep quality predictions by extracting randomly the positive and negative labels from a uniform distribution. The second baseline, denominated "a-priori", always predict a constant value, chosen as the majority class, which in this case is the positive class. We adopt the *balanced accuracy* as the evaluation metric for the experiments, given the imbalance in the class distributions when testing [9]. We compare the performance of these baseline classifiers with the other models using the Wilcoxon signed-rank statistical significance test [18,21] with a threshold of 0.05.

Classification Results. For the *Leave One Participant Out* cross validation, we show results in Table 2. While small variations in performance are present across window size and device, all classifiers do not achieve accuracies higher than 0.65, with most models not higher then the baselines (0.5 RG and 0.6 a-priori). Indeed, only one model (AdaBoost, whole night, E4) achieves a balanced accuracy higher than 0.6. However, all models are not statistically different (p-value threshold $\alpha = 0.05$) from the a-priori baseline. The results suggest that both interpersonal variability and the limited number of participants do not allow to achieve significant performance, with respect to the baselines, when testing on an unseen participant [2].

For the *Leave Out Session Out* cross validation paradigm, we report results in Table 3. From these, we see that models trained on the whole night achieve a lower performance than models trained over smaller windows. The results show that for the 5 min window most of the models are able to recognize the sleep quality for an already existing participant, surpassing the baselines (>0.62). A model trained on the 60 min window, using Oura features, has the highest overall average accuracy (0.76). For the whole night, one of the models can reach a performance of 0.66 using the E4 features. Accordingly, the performance

Table 2. Average balanced accuracy, with standard errors, for three devices in sleep quality recognition task using different window size and *Leave One Participant Out* cross validation. The a-priori baseline always predicts the positive class, while the Random Guess (RG) uses a uniform distribution to make predictions

Window size Device/ Model	5 mins		10 mins		60 mins		whole night	
	Oura	E4	Oura	E4	Oura	E4	Oura	E4
DT	0.53 ± 0.13	0.45 ± 0.05	0.46 ± 0.03	0.50 ± 0.09	0.44 ± 0.05	0.43 ± 0.09	0.42 ± 0.07	0.48 ± 0.03
NB	0.49 ± 0.01	0.50 ± 0.00	0.47 ± 0.03	0.49 ± 0.01	0.42 ± 0.05	0.47 ± 0.03	0.45 ± 0.02	0.47 ± 0.03
SVM	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.51 ± 0.01	0.45 ± 0.05	0.50 ± 0.00	0.5 ± 0.03	0.50 ± 0.00
MLP	0.50 ± 0.00	0.47 ± 0.03	0.50 ± 0.00	0.49 ± 0.02	**0.55 ± 0.05**	**0.52 ± 0.02**	0.42 ± 0.06	0.54 ± 0.07
RF	0.46 ± 0.04	0.49 ± 0.01	0.52 ± 0.03	**0.55 ± 0.07**	0.42 ± 0.06	0.41 ± 0.06	0.36 ± 0.06	0.43 ± 0.05
XGBoost	**0.55 ± 0.02**	0.48 ± 0.02	0.52 ± 0.03	0.50 ± 0.04	0.52 ± 0.02	0.52 ± 0.03	0.38 ± 0.07	0.5 ± 0.05
AdaBoost	0.48 ± 0.06	0.49 ± 0.04	**0.53 ± 0.05**	0.50 ± 0.03	0.53 ± 0.06	0.51 ± 0.05	0.43 ± 0.05	**0.64 ± 0.05**
QDA	0.54 ± 0.06	0.5 ± 0.00	0.49 ± 0.01	0.45 ± 0.06	0.49 ± 0.01	0.51 ± 0.04	**0.51 ± 0.07**	0.55 ± 0.04
KNN	0.53 ± 0.04	**0.53 ± 0.05**	0.44 ± 0.04	0.51 ± 0.03	0.50 ± 0.07	0.45 ± 0.03	0.41 ± 0.1	0.45 ± 0.03
GP	0.48 ± 0.02	0.49 ± 0.01	0.5 ± 0.01	0.45 ± 0.04	0.50 ± 0.04	0.49 ± 0.11	0.45 ± 0.1	0.47 ± 0.1
RG	0.39 ± 0.10		0.62 ± 0.04		0.44 ± 0.09		0.62 ± 0.06	
a-priori	0.50 ± 0.0		0.50 ± 0.0		0.50 ± 0.0		0.5 ± 0.0	

of the models diminishes when using the whole sleep session, compared to smaller window sizes. The results also suggest that there is no real advantage between models trained with data from Oura ring or E4 wristband, with all window experiments achieving accuracies higher then 0.7 with at least one model (best baseline 0.61). From the experiments, we can conclude that both devices achieve comparable performance in the sleep quality recognition task. With our results in Sect. 4 and Subsect. 5.2, these classification task supports that the devices used are interchangeable with respect to heart rate data. We also find that it is better to adopt a windowed data, as opposed to using the whole night session, when performing sleep quality recognition. However, it is worth noting how, given the limited amount of data, the standard errors evaluated are quite large. This means that no result is statistically significant with respect to the a-priori baseline, increasing the available data would allow to mitigate this problem.

6 Limitations and Future Work

The main limitation of our work is the small number of participants in the dataset (five). We also only collected an average of 16 nights per participants. In future work, performing a data collection with more participants and for more nights could lead to further insights. This is especially true for the Polar chestbelt, since we do not use in the subjective sleep recognition task given the limited number (18) of sleep sessions collected with this device. The use of more HR tracking wearable devices could be explored. As suggested by [2], the use of a subjective sleep quality score can also hinder a machine learning task, as such we are considering exploring the devices performance compared to an additional objective measure.

Table 3. Average balanced accuracy, with standard errors, for three devices in sleep quality recognition task using different window size and *Leave One Session Out* cross validation. The a-priori baseline always predicts the positive class, while the Random Guess (RG) uses a uniform distribution to make predictions.

Window size Device/ Model	5 mins		10 mins		60 mins		whole night	
	Oura	E4	Oura	E4	Oura	E4	Oura	E4
DT	0.62 ± 0.05	0.71 ± 0.05	0.71 ± 0.05	0.71 ± .05	0.74 ± 0.05	0.68 ± 0.05	0.53 ± 0.05	0.60 ± 0.05
NB	0.71 ± 0.05	0.68 ± 0.05	0.71 ± 0.05	0.68 ± 0.05	0.60 ± 0.06	0.66 ± 0.05	0.54 ± 0.05	0.63 ± 0.05
SVM	0.50 ± 0.06	0.53 ± 0.06	0.42 ± 0.06	0.53 ± 0.06	0.51 ± 0.06	0.44 ± 0.06	0.52 ± 0.05	0.55 ± 0.05
MLP	0.57 ± 0.06	0.64 ± 0.05	0.56 ± 0.06	0.59 ± 0.06	0.59 ± 0.06	**0.72 ± 0.05**	0.55 ± 0.05	0.65 ± 0.05
RF	0.69 ± 0.05	**0.72 ± 0.05**	**0.72 ± 0.05**	**0.74 ± 0.05**	0.74 ± 0.05	0.65 ± 0.05	0.61 ± 0.05	**0.66 ± 0.05**
XGBoost	0.66 ± 0.05	**0.72 ± 0.05**	**0.72 ± 0.05**	0.72 ± 0.05	**0.76 ± 0.05**	0.65 ± 0.05	0.60 ± 0.05	0.63 ± 0.05
AdaBoost	**0.72 ± 0.05**	**0.72 ± 0.05**	**0.72 ± 0.05**	0.72 ± 0.05	0.64 ± 0.05	**0.72 ± 0.05**	0.51 ± 0.05	0.63 ± 0.05
QDA	0.64 ± 0.05	0.64 ± 0.05	0.59 ± 0.06	0.62 ± 0.05	0.50 ± 0.06	0.64 ± 0.05	0.50 ± 0.05	0.52 ± 0.05
KNN	0.66 ± 0.05	0.70 ± 0.05	0.71 ± 0.05	0.68 ± 0.05	0.66 ± 0.05	0.65 ± 0.05	0.52 ± 0.05	0.59 ± 0.05
GP	0.70 ± 0.05	0.72 ± 0.05	0.71 ± 0.05	0.71 ± 0.05	0.70 ± 0.05	0.55 ± 0.06	**0.64 ± 0.05**	0.59 ± 0.05
RG	0.61 ± 0.05		0.57 ± 0.06		0.39 ± 0.05		0.41 ± 0.05	
a-priori	0.61 ± 0.05		0.61 ± 0.06		0.61 ± 0.05		0.59 ± 0.05	

7 Conclusion

We run a data collection campaign for 30 nights to collect HR data during sleep using Oura ring, Empatica E4 wristband and Polar chestbelt, in the wild, along with self reports about sleep behaviour. We provide the dataset to other researchers upon request to extend our data analysis. Then, we investigate the interchangeability of HR data collected from these wearables. To this goal, we run an extensive data analysis. We find that there is a high positive correlation between the HR data from the three devices based on Spearman's correlation coefficient. Using bias analysis, we also estimate that the Oura ring's HR signal has less variations with respect to the ECG-based Polar chestbelt, compared to data from the E4 wristband. We also assess the difference between time-domain features extracted from the three devices for different windows sizes, finding them negligible or small in most cases. Finally in order to evaluate the impact of such small differences, we employ these features in a machine learning task to predict subjective sleep quality. We find that, when testing on a new sleep session, there is not appreciable difference between models trained on features extracted from Oura ring's or E4 wristband's HR signals. We also find that a higher performance is achieved when separating the sleep session into non-overlapping windows, as opposed to using the whole night's data. In conclusion, our results suggest interchangeability among the devices. Even with the outlined limitations of our study, we believe that the three devices can be used in broader settings, e.g., health tracking, with similar outcomes.

Acknowledgement. This contribution is supported by the Swiss National Science Foundation (SNSF) through the grant 205121 _197242 for the project "PROSELF: Semi-automated Self-Tracking Systems to Improve Personal Productivity". Shkurta Gashi is supported by an ETH AI Center postdoctoral fellowship.

References

1. Alchieri, L., et al.: On the impact of lateralization in physiological signals from wearable sensors (2022)
2. Alecci, L., et al.: On the mismatch between measured and perceived sleep quality. In: Proceedings of the 2022 UbiComp (2022). https://doi.org/10.1145/3544793.3563412
3. Altini, M., et al.: The promise of sleep: a multi-sensor approach for accurate sleep stage detection using the oura ring. Sensors **21**(13) (2021)
4. Armstrong, R.A.: When to use the B onferroni correction. Ophthalmic Physiol. Opt. **34**(5) (2014)
5. Assaf, M., Rizzotti-Kaddouri, A., Punceva, M.: Sleep detection using physiological signals from a wearable device. In: Inácio, P.R.M., Duarte, A., Fazendeiro, P., Pombo, N. (eds.) HealthyIoT 2018. EICC, pp. 23–37. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-30335-8_3
6. Barika, R., et al.: A smart sleep apnea detection service. In: 17th International Conference on CM. The British Institute of NDT (2021)
7. Bland, J.M., et al.: Measuring agreement in method comparison studies. Stat. Methods Med. Res. **8**(2) (1999)
8. Breiman, L.: Random forests. Mach. Learn. **45**(1) (2001)
9. Brodersen, K.H., et al.: The balanced accuracy and its posterior distribution. In: 20th ICPR. IEEE (2010)
10. Buysse, D.J., et al.: The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatry Res. **28**(2), 193–213 (1989)
11. Cakmak, A.S., et al.: An unbiased, efficient sleep-wake detection algorithm for a population with sleep disorders: change point decoder. Sleep **43**(8) (2020)
12. Carlozzi, N.E., et al.: Daily variation in sleep quality is associated with health-related quality of life in people with spinal cord injury. Arch. Phys. Med. Rehabil. **103**(2) (2022)
13. Chawla, N.V., et al.: Smote: synthetic minority over-sampling technique. JAIR **16** (2002)
14. Chee, N.I., et al.: Multi-night validation of a sleep tracking ring in adolescents compared with a research actigraph and polysomnography. Nat. Sci. Sleep **13** (2021)
15. Chinoy, E.D., et al.: Performance of four commercial wearable sleep-tracking devices tested under unrestricted conditions at home in healthy young adults. Nat. Sci. Sleep **14** (2022)
16. Cliff, N.: Dominance statistics: ordinal analyses to answer ordinal questions. Psychol. Bull. **114**(3), 494 (1993)
17. Cole, C.R., Blackstone, E.H., Pashkow, F.J., Snader, C.E., Lauer, M.S.: Heart-rate recovery immediately after exercise as a predictor of mortality. N. Engl. J. Med. **341**(18), 1351–1357 (1999)
18. Conover, W.J.: Practical Nonparametric Statistics, vol. 350. Wiley, Hoboken (1999)
19. Cortes, C., et al.: Support-vector networks. Mach. Learn. **20**(3) (1995)
20. Duda, R.O., et al.: Pattern Classification and Scene Analysis, vol. 3. Wiley, New York (1973)
21. Field, A., et al.: How to Design and Report Experiments. Sage (2002)
22. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)

23. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stati. (2001)
24. Gardner, M.W., et al.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos. Environ. **32**(14–15) (1998)
25. Gashi, S., et al.: Using unobtrusive wearable sensors to measure the physiological synchrony between presenters and audience members. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **3**(1), 1–19 (2019)
26. Gashi, S., et al.: The role of model personalization for sleep stage and sleep quality recognition using wearables. IEEE Pervasive Comput. **21**, 69–77 (2022)
27. Ghorbani, S., et al.: Multi-night at-home evaluation of improved sleep detection and classification with a memory-enhanced consumer sleep tracker. Nat. Sci. Sleep **14** (2022)
28. Gilgen-Ammann, R., et al.: RR interval signal quality of a heart rate monitor and an ECG Holter at rest and during exercise. EJAP **119** (2019)
29. Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, vol. 2. Springer, Heidelberg (2009)
30. Hellhammer, J., et al.: The physiological response to trier social stress test relates to subjective measures of stress during but not before or after the test. Psychoneuroendocrinology **37**(1), 119–124 (2012)
31. Hernandez, J., Morris, R.R., Picard, R.W.: Call center stress recognition with person-specific models. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011. LNCS, vol. 6974, pp. 125–134. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24600-5_16
32. Imtiaz, S.A.: A systematic review of sensing technologies for wearable sleep staging. Sensors **21**(5) (2021)
33. Joshi, A., et al.: Likert scale: explored and explained. Br. J. Appl. Sci. Technol. **7**(4) (2015)
34. Kelleher, J.D., Mac Namee, B., D'arcy, A.: Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press (2020)
35. Kendall, M.G., et al.: The Advanced Theory of Statistics. The Advanced Theory of Statistics, 2nd edn (1946)
36. Kromrey, J.D., et al.: Analysis options for testing group differences on ordered categorical variables: an empirical investigation of type I error control and statistical power. MLRV **25**(1) (1998)
37. Mehrabadi, M.A., et al.: Sleep tracking of a commercially available smart ring and smartwatch against medical-grade actigraphy in everyday settings: instrument validation study. JMIR mHealth uHealth **8**(11) (2020)
38. Miller, D.J., et al.: A validation study of a commercial wearable device to automatically detect and estimate sleep. Biosensors **11**(6) (2021)
39. Min, J.K., Doryab, A., Wiese, J., Amini, S., Zimmerman, J., Hong, J.I.: Toss'n'turn: smartphone as sleep and sleep quality detector. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 477–486 (2014)
40. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. JMLR **12** (2011)
41. Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
42. Raskovic, D., et al.: Medical monitoring applications for wearable computing. Comput. J. **47**(4), 495–504 (2004)

43. Reinhardt, T., et al.: Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the Mannheim Multicomponent Stress Test (MMST). Psychiatry Res. **198**(1), 106–111 (2012)
44. Roberts, D.M., et al.: Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography. Sleep **43**(7) (2020)
45. Sano, A., et al.: Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In: Proceedings of the IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN 2015). IEEE (2015)
46. Sano, A., et al.: Multimodal ambulatory sleep detection using LSTM recurrent neural networks. IEEE J. Biomed. Health Inform. **23**(4), 1607–1617 (2019)
47. Schmidt, P., Reiss, A., Dürichen, R., Van Laerhoven, K.: Wearable-based affect recognition—a review. Sensors **19**(19), 4079 (2019)
48. Scott, H., et al.: The development and accuracy of the THIM wearable device for estimating sleep and wakefulness. Nat. Sci. Sleep **13** (2021)
49. Siirtola, P., et al.: Using sleep time data from wearable sensors for early detection of migraine attacks. Sensors **18**(5) (2018)
50. Stone, J.D., et al.: Evaluations of commercial sleep technologies for objective monitoring during routine sleeping conditions. Nat. Sci. Sleep **12** (2020)
51. Swain, P.H., et al.: The decision tree classifier: design and potential. IEEE Trans. Geosci. Electron. **15**(3) (1977)
52. Taylor, S.A., et al.: Personalized multitask learning for predicting tomorrow's mood, stress, and health. IEEE Trans. Affect. Comput. **11**, 200–213 (2017)
53. Williams, C.K., et al.: Gaussian Processes for Machine Learning, vol. 2. MIT Press, Cambridge (2006)
54. Yan, S., et al.: Estimating individualized daily self-reported affect with wearable sensors. In: 2019 IEEE ICHI (2019)

A Low-Cost Wearable System to Support Upper Limb Rehabilitation in Resource-Constrained Settings

Md. Sabbir Ahmed[1](✉), Shajnush Amir[1], Samuelson Atiba[2], Rahat Jahangir Rony[1], Nervo Verdezoto Dias[2], Valerie Sparkes[3], Katarzyna Stawarz[2], and Nova Ahmed[1]

[1] Design Inclusion and Access Lab (DIAL), North South University, Dhaka, Bangladesh
msg2sabbir@gmail.com, {shajnush.amir,rahat.rony, nova.ahmed}@northsouth.edu
[2] School of Computer Science and Informatics, Cardiff University, Cardiff, UK
{atibas,verdezotodiasn,stawarzk}@cardiff.ac.uk
[3] School of Healthcare Sciences, Cardiff University, Cardiff, UK
sparkesv@cardiff.ac.uk

Abstract. There is a lack of professional rehabilitation therapists and facilities in low-resource settings such as Bangladesh. In particular, the restrictively high costs of rehabilitative therapy have prompted a search for alternatives to traditional in-patient/out-patient hospital rehabilitation moving therapy outside healthcare settings. Considering the potential for home-based rehabilitation, we implemented a low-cost wearable system for 5 basic exercises namely, *hand raised, wrist flexion, wrist extension, wrist pronation, and wrist supination*, of upper limb (UL) rehabilitation through the incorporation of physiotherapists' perspectives. As a proof of concept, we collected data through our system from 10 Bangladeshi participants: 9 researchers and 1 undergoing physical therapy. Leveraging the system's sensed data, we developed a diverse set of machine learning models. And selected important features through three feature selection approaches: filter, wrapper, and embedded. We find that the Multilayer Perceptron classification model, which was developed by the embedded method Random Forest selected features, can identify the five exercises with a ROC-AUC score of 98.2% and sensitivity of 98%. Our system has the potential for providing real-time insights regarding the precision of the exercises which can facilitate home-based UL rehabilitation in resource-constrained settings.

Keywords: Upper limb rehabilitation · Low-resource · Wearable · Machine learning · Exercises · Physiotherapy · Bangladesh · Digital health · Low-cost wearable

---

Md. S. Ahmed and S. Amir—Equal contribution.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 33–45, 2023.
https://doi.org/10.1007/978-3-031-34586-9_3

1 Introduction

Upper limb impairment, a reduction or loss of limb function, is one of the most common consequences of acquired brain injury (ABI) [3]. In Bangladesh, ABI due to stroke and trauma is the leading cause of death and disability, representing an immense economic cost to the nation [20]. Over 97% of people with an ABI are diagnosed with some form of limb weakness that affects their ability to independently perform daily activities [3, 20]. In addition, there are very few care facilities and professionals [16, 20] limiting access to rehabilitation services, which increases the risk of long-term disability [20]. The high costs associated with hospital-based therapy continue to be a major barrier [20]. For example, in 2016, the typical Bangladeshi household income per month was 15,988 BDT ($189.76) [2] while the monthly cost for hospital-based rehabilitation in 2017 was 27,852 BDT ($328) [20]. High costs force one to choose between poverty and lifelong disability.

These challenges have created an opportunity for the use of technology to support home-based rehabilitation, especially in remote areas of Bangladesh. Technology-based rehabilitation in the home offers greater accessibility and convenience in relation to the time spent attending face-to-face appointments, thus reducing the overall costs of rehabilitation [17, 19]. Several technologies have been deployed for use in upper limb (UL) rehabilitation, including rehabilitation robots which actively assist patients to perform rehabilitation exercises [7], electrical stimulation which uses an electrical current to stimulate muscles in the affected limb [24], and wearable sensor devices which capture the patient's movements during rehabilitation exercises [17]. However, the robots are often large and expensive [19], and the electrical stimulation hardware requires expert knowledge and dexterous manipulation to set up [24]. Though there are several low-cost rehabilitation systems, there is a lack of computational models (e.g., [1, 11]) that could enable the systems to automatically identify the exercises and provide feedback to the patients and caregivers in real-time.

Therefore, we present a low-cost wearable system that incorporates machine learning models to support UL rehabilitation. Our contribution is twofold:

- We present a low-cost (around $16) system for recording and monitoring exercises to support UL rehabilitation.
- We develop machine learning (ML) models based on 14 algorithms and three feature selection (FS) approaches and show that the Multilayer Perceptron (MLP) model performed best with a ROC-AUC and precision score of 98.2%.

Overall, our system can facilitate home-based UL rehabilitation and real-time monitoring of the patients in low-resource settings.

2 Related Work

2.1 Approaches to Upper Limb Rehabilitation

Conventional rehabilitation is typically conducted in a controlled hospital environment. The methods for UL rehabilitation include mental imagery and action observation [12], constraint-induced movement therapy [14], and task-specific training [10]. Hospital-based task-specific training can lead to improved rehabilitation outcomes when administered frequently over an extended period [4, 10]. However, trained competencies acquired in the hospital environment, such as grasping and reaching, often fail to transfer to home and work environments, since trained movements may not correspond to activities in daily life [9].

Compared to hospital-based rehabilitation, home-based rehabilitation focused on everyday actions has been shown to achieve significantly better outcomes with regard to training transfer. This is because the training exercises are carried out within the relevant context where they would occur daily [21]. In addition, since home-based rehabilitation reduces the need for frequent hospital visits, and does not require expensive facilities, the cost of rehabilitation can be greatly lowered.

2.2 Technologies to Support Upper Limb Rehabilitation

In response to the demand for technological interventions for rehabilitation, several technologies have been deployed to support UL rehabilitation. These technologies include rehabilitation robots [7], electrical stimulation [24], and wearable sensor devices [17]. Rehabilitation robots, such as exoskeletons and soft wearable robots [7, 19], allow for precise movement control while providing assistance to weakened or paralyzed limbs during rehabilitation. But they are very expensive, not easily portable, and hard to wear and undress. Also, they often pose a safety risk when there is a misalignment between the robot and the human anatomy [19]. Consequently, rehabilitation robots are deployed in controlled hospital environments where the expertise is available to support clinical and rehabilitation practices. Thus, these technologies are often unsuitable for home-based rehabilitation.

Electrical stimulation (ES) for rehabilitation is focused on producing motor responses in muscles that are weakened or paralyzed due to an upper motor neuron injury, as is the case in people with ABI [24]. Its major setback in rehabilitation is the possibility of fatigue due to neurotransmitter depletion or propagation failure. When such fatigue sets in, the muscle fibers are not sufficiently stimulated and hence do not gain strength [24]. Therefore, expert knowledge is required to control the parameters of the stimulation provided. The need for expert monitoring and expensive specialized equipment restricts the deployment of ES in home-based rehabilitation settings.

Wearable devices are a widely explored system for UL rehabilitation [25]. They are lightweight, easy to put on and take off, cost-effective, and easy to operate [1, 17]. In addition, it is feasible to deploy them in the home as an alternative and/or complement to hospital-based rehabilitation [5]. Due to their low cost, they are also suitable for low-income settings. As such, researchers developed systems for the Global South focusing on exercises such as flexion, extension, abduction, horizontal abduction [1], supine [11], etc. However, they rely on visualization techniques which may not be precise enough to account for subtle differences to accurately identify UL exercises.

3 System Development

3.1 Understanding Physiotherapists' Perspectives

The developed system was informed by interviews with four Bangladeshi (3 men, 1 woman) physiotherapists, who helped us to identify basic exercises that were important for UL rehabilitation. To provide context, we summarize the key points that informed the design of our system; detailed interview results are reported elsewhere.

In Bangladesh, physiotherapy focuses on basic movements aiming to strengthen the muscles. However, the limited access to treatment was further worsened during the pandemic, as many centers closed down and physiotherapy sessions were discontinued:

"All patients' treatments do not complete at-home services. Sometimes there are required machines. So, that is not possible at home rather than in centers. In Bangladesh, good physio centers do not have many branches, so that people can't get support during COVID-19."- Physiotherapist 1.

This situation highlights the need for home-based physiotherapy. However, physiotherapists mentioned issues with the accuracy of movements when practicing at home, which can have a negative impact on patients:

"For hand movements, patients sometimes lift the wrong shoulder. Here movement is done but wrong. Detecting accurate movement is necessary" - Physiotherapist 4.

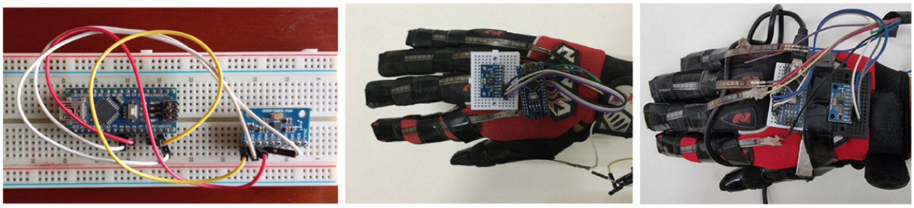
Maintaining accuracy at home requires a system that can monitor the patients' exercises. They have suggested key 5 exercises necessary for UL progress:

- *Hand raised*: It is an exercise where the hands are kept up 90° and the shoulders are kept straight.
- *Wrist flexion*: It is the bending of the hand down at the wrist where the palm faces toward the arm.
- *Wrist extension*: It is the opposite of flexion where the movement of the hand is backward, towards the forearm's posterior side.
- *Wrist pronation*: In pronation, the forearm or palm faces down.
- *Wrist supination*: In this exercise, the forearm or palm faces up.

3.2 System Design

To develop a low-cost system that can facilitate the identification of the aforementioned exercises unobtrusively, we used Arduino Nano (price ~$7) and inertial measurement unit (IMU) sensor MPU-9250 9-DOF (price ~$9) where the IMU consists of an accelerometer, gyroscope, and magnetometer. Firstly, a basic prototype (Fig. 1(a)) was developed to ensure the component level accuracy, which was followed by a working prototype (Fig. 1(b)) on a glove that had a flex sensor placed on each finger. However, the sensors' placement added extra noise which was finally modified (Fig. 1(c)) by keeping the sensors further from the finger.

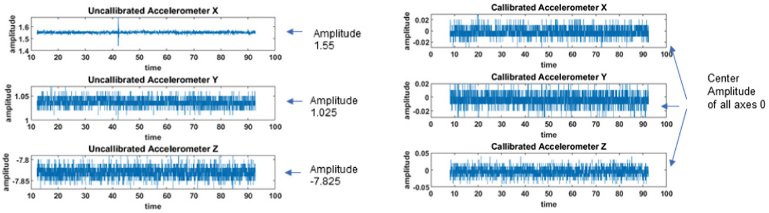
Data Recording Application: A data recorder was developed to prompt the user to record the data of each exercise. The user could control the beginning of the recording



(a) Basic prototype (b) Working prototype (c) Final design

Fig. 1. Development of the system.

of each exercise. In the data recorder, the delay between each movement of the same exercise can be modified as the users may have different preferences.



(a) Uncalibrated data (b) Calibrated data

Fig. 2. (a) Uncalibrated and (b) calibrated data of the accelerometer.

Initialization and Stabilization: The sampling rate of the system was 90 Hz. The raw data of the accelerometer, as an example, portrayed in Fig. 2(a), is taken before any hand gesture takes place to get a sense of the IMU sensor's data. As seen in Fig. 2(a), the starting amplitude without any hand gestures is non-zero. The uncalibrated values are due to gravity, the earth's magnetic flux coupled with the sensor's offset values which implies that the sensors' starting value will be different in different starting positions of the sensor. To ensure amplitude values are stabilized, a calibration routine was implemented where during the initialization of the wearable at a flat surface, 3000 readings are taken and averaged. This averaged value is subtracted from all subsequent sensor data to remove the offset. The data after the calibration is illustrated in Fig. 2(b). Due to the calibration routine, the amplitude of all 3 axes of data points is always set to zero at the initiation of the device for rest position.

Precision Adjustment: Initially, there was low precision as the data recorder registered integer values to reduce processing time. Due to the absence of decimal points, the data plots were not continuous (Fig. 3). To get the decimal points, we multiplied the sensor data by 100,000 and divided the integer data by that value, thereby incorporating the lost decimal values and increasing the precision. E.g., for the low precision gyroscope data, there is a staircase effect, but for the high precision data, the plot line is continuous as decimal points are incorporated (Fig. 3).

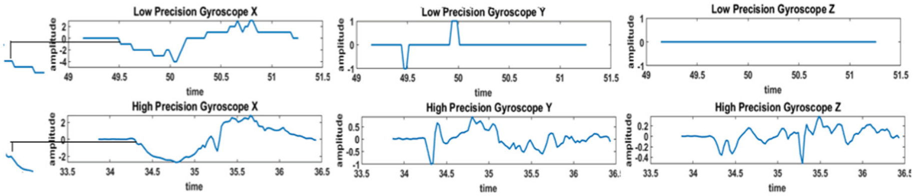


Fig. 3. Improved data precision of the gyroscope.

3.3 Validation of the Developed System

For validation, our system's retrieved data was compared with a Samsung Galaxy S6 smartphone (Table 1) which has been found as a promising device to sense data [18]. To get the sensors' data from Galaxy S6, we used the Physics *Toolbox Sensor Suite Pro* app which is available in the Play Store. We recruited 2 participants and each participant performed each exercise through our system and also through Galaxy S6. During validation, both the prototype and the smartphone were connected to the user at the same time and the data was captured in both devices simultaneously. Therefore, both sets of data captured the same exercise movement data.

Table 1. Comparison of our system's retrieved data with the Galaxy S6.

Axis	Accelerometer			Gyroscope			Magnetometer		
	Peak diff	Noise variance		Peak diff	Noise variance		Peak diff.	Noise variance	
		Our system	Galaxy S6		Our system	Galaxy S6		Our system	Galaxy S6
X	8.98%	$5.7 * 10^{-7}$	0.0942	1.14%	$4.58 * 10^{-5}$	$9.42 * 10^{-5}$	1.14%	$4.58 * 10^{-5}$	$9.42 * 10^{-5}$
Y	5.98%	$3.76 * 10^{-7}$	$2.07 * 10^{-4}$	0.58%	$6.11 * 10^{-8}$	$1.05 * 10^{-4}$	0.58%	$6.12 * 10^{-8}$	$1.05 * 10^{-4}$
Z	1.22%	$8.12 * 10^{-5}$	$8.19 * 10^{-4}$	16.17%	$2.47 * 10^{-7}$	$1.24 * 10^{-5}$	16.17%	$2.47 * 10^{-7}$	$1.24 * 10^{-5}$

For comparison, we calculated the noise variance using the MATLAB function *evar* [8] which estimates the variance of additive noise. Peak difference was calculated using the formula $peak\_dif = \frac{|(peak\_value_{our\_syetm} - peak\_value_{S6})|}{peak\_value_{S6}}$. In the Z-axis of the accelerometer and the X-axis of the gyroscope and magnetometer, the peak difference between our system and S6 retrieved data was less than 1.5% (Table 1). Though in some cases the peak difference was around 16%, our system had much lower noise variance. Thus, our system's data was comparable to the S6 phone with better noise performance.

4 Methodology

4.1 Participants and Research Ethics

As a proof of concept, we conducted a study in Bangladesh with 10 participants, nine researchers, and one undergoing physiotherapy. On average, each participant provided data on 10.8 different days (SD = 7.20, Minimum = 3, Maximum = 30, Median = 10), and on each day, each participant did each exercise 5 times. While collecting data, we labeled the 5 exercises (e.g., wrist flexion) so that the ML models' prediction could be evaluated.

The study was approved by the North South University IRB/ERC committee (2020/OR-NSU/IRB-No.0501). We received the participants' signed consent forms.

4.2 ML Model Development

4.2.1 Feature Extraction and Selection

For each participant, we calculated 6 types of data, namely, mean, standard deviation (SD), interquartile range (IQR), skewness, kurtosis, and entropy over the time periods based on the accelerometer, magnetometer, and gyroscope sensed data from each of the 3 axes (X, Y, Z) separately. In total, we extracted 54 features (6 types of data * 3 sensors * 3 axes) from each participant. But 36 (67%) of the features' data were not normally distributed, and thus, we normalized the data instead of standardization.

In general, feature selection (FS) methods can be grouped into 3 categories [15]: wrapper, filter, and embedded method. As a wrapper method, we used the Boruta algorithm which is an all-relevant FS approach [23]. We tuned the maximum depth of Boruta's base estimator Random Forest (RF) algorithm and the range was 3 to 7 which is suggested to use [6]. We used the Information Gain (IG) and RF algorithms as the filter and embedded methods respectively. IG and RF algorithms work by a minimal-optimal method whereas, unlike the all-relevant FS approach, it does not inform a fixed set of features to be used. Therefore, for the IG and RF methods, we used the maximum length of the Boruta selected features set as the upper boundary and 1 as the lower boundary of the number of features to be selected.

4.2.2 Model Development and Validation

Based on the "No Free Lunch" theorem, there is no algorithm that can perform best for all problems. Hence, we developed models (Fig. 4) by exploring a diverse set of ML algorithms: Logistic Regression (Logit), K-Nearest Neighbor (KNN), Support Vector Classifier (SVC), Gaussian Naïve Bayes (GNB), Decision Trees, Random Forest (RF), Gradient Boosting (GB), Light GBM, AdaBoost, Extra Tree, CatBoost, Extreme Gradient Boosting, and Multilayer Perceptron (MLP). In addition, a Dummy classifier was used as the baseline which predicts regardless of the input features. Inspired by Vabalas et al. [22], we used the nested cross-validation approach which shows generalizable performance. In the outer loop, there was Leave One Out Cross Validation (LOOCV) where we divided the dataset into n equal portions where each portion presenting a participant's data. Then, we used $n - 1$ participants' data to select the best set of features, and in the inner loop, to tune the hyper-parameters, we used a 5-fold CV maximizing the macro-F1 score. For tuning, we used the Bayesian search technique. After finding the best estimator, we predicted the class of the left participant's exercises who was not involved in FS and hyper-parameter tuning stages. We repeated this process to predict the exercise class of each of the 10 participants.

We evaluated the models' performance by comparing the labeled class with the models' predicted exercise and calculated the precision, sensitivity, F1, and ROC-AUC (Area Under the Receiver Operating Characteristic Curve). Each evaluation metric's

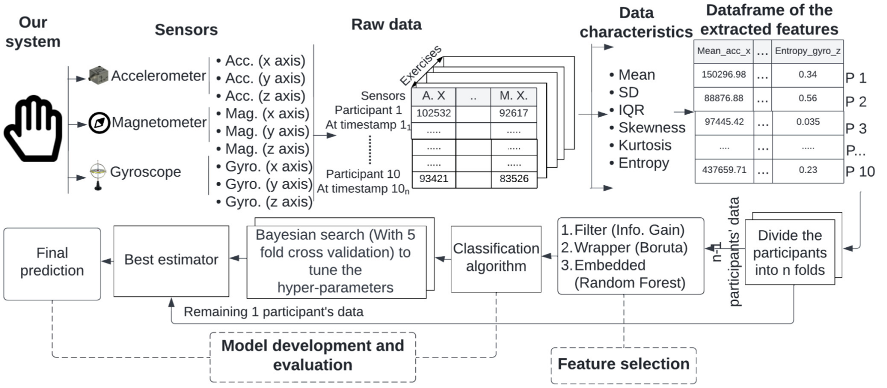
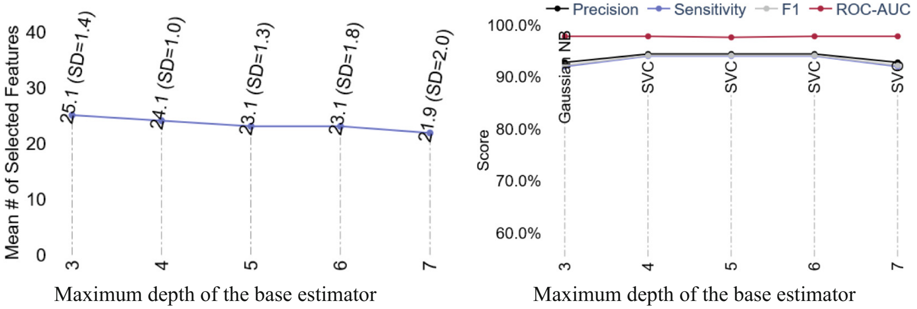


Fig. 4. ML pipeline to identify each exercise.

score was macro-averaged by calculating the simple arithmetic mean of all the 5 class scores of the evaluation metric (e.g., precision). It should be noted that in our study, each participant performed each of the 5 exercises, which means there is no class imbalance. In addition, to make the models unbiased, none of the 5 exercise data of the test participants were used in the training phase.

5 Results

5.1 Predicting the Exercises



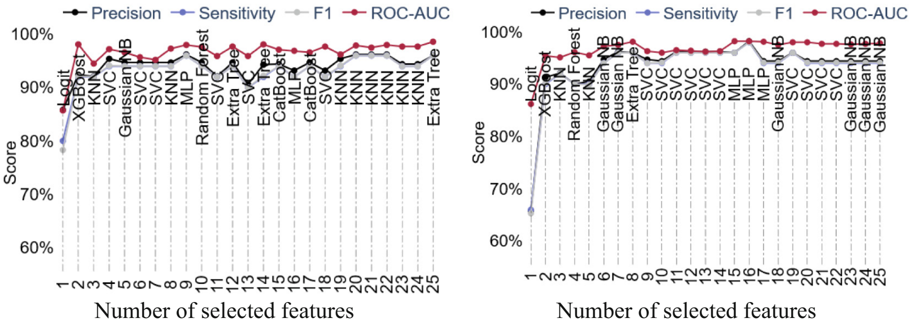
(a) Number of selected features (b) Best models' performance

Fig. 5. (a) Number of selected features and (b) performance of the best model while selecting the features by tuning the maximum depth of the base estimator of Boruta.

To identify each exercise from the sensor retrieved data, we tuned the maximum depth of the base estimator Random Forest (RF) in the all-relevant FS approach Boruta. The number of selected features was lower with the increase in the maximum depth (Fig. 5(a)). It is apparent that at maximum depths 4, 5, and 6, the best performing model SVC has almost identical performance (Precision = 94.4%, F1 = 94.1%, Sensitivity

= 94%) where SVC can identify 94% of exercises accurately (Fig. 5(b)). However, the mean number of selected features in each iteration of LOOCV is 24.1 (SD = 1.0) at depth 4 whereas it is 23.1 at depths 5 and 6 (Fig. 5(a)). As the ROC-AUC score is 97.6% and 97.8% at depths 5 and 6 respectively, we consider SVC at depth 6 as the best model due to having relatively higher predictability.

In the Boruta FS approach, on average, the maximum number of selected features was 25.1 which is at depth 3 (Fig. 5(a)). Therefore, as discussed in Sect. 4.2.1, we set 25 as the upper boundary of the number of features to be selected in the filter and embedded FS methods. In the filter method Information Gain (IG), when there is only 1 feature selected, the Logit model performed best with an F1 score of around 80% (Fig. 6(a)). However, the MLP model based on 9 important features in each iteration of LOOCV had a maximum F1 of 96% and a ROC-AUC score of 98%. Though at features 9, 20, 21, 22, and 25 the performance of the best model is almost similar, the model based on 9 features were selected as best due to having lower features.



(a) IG (b) RF

Fig. 6. Best models' performance when a number of important features are selected through the (a) filter method IG and (b) embedded method RF algorithm.

In the embedded method RF selected 9 important features, the best model SVC had a precision of 94.7%, an F1 score of 94%, and a ROC-AUC score of 96% (Fig. 6(b)) which was lower than the performance of the best model based on IG selected 9 features. However, the MLP model based on the RF selected 16 features in each iteration of LOOCV has a ROC-AUC score of 98.2%, precision of 98.2%, and F1 score of 98%. This MLP model had higher performance than any other models based on the Boruta (Fig. 5) and IG selected features (Fig. 6(a)).

Though we developed models based on 14 algorithms, we found conventional ML models' (e.g., SVC, GN) superior performance. In the Boruta selected features, the best model in each depth was either SVC or the GNB (Fig. 5(b)). Also, in IG (Fig. 6 (a)) and RF (Fig. 6(b)) selected feature-based models, the best performing model in most cases was KNN and SVC. Apart from these, though there was a single tree-based model among the top-5 models in the case of each FS method's best set of features, we found 3 conventional algorithm-based models (Table 2) which shows their robustness to identify

Table 2. Performance of the top-5 classifiers and baseline Dummy classifier, based on the best (in terms of ML models' performance) set of features of each FS method. "# of features" present the number of features used in each iteration of LOPOCV. E: Extra.

Filter method IG (# of features = 9)					Wrapper method Boruta (average # of features = 23.1 (SD: 1.8))					Embedded method RF (# of features = 16)				
Model Name	Precision	Sensitivity	F1	ROC AUC	Model Name	Precision	Sensitivity	F1	ROC AUC	Model Name	Precision	Sensitivity	F1	ROC AUC
MLP	96.2	96	96	98	SVC	94.4	94	94.1	97.8	MLP	98.2	98	98	98.2
SVC	93.2	92	92	96.8	GNB	92.8	92	92.2	97.8	GNB	94.4	94	94.1	97.8
Logit	90	90	90	96	MLP	91	90	90.1	95.9	SVC	94	94	93.9	98
E. Tree	90	90	90	95.9	E. Tree	90.4	90	90.1	98.3	KNN	92.3	90	89.7	93.8
GNB	90.1	90	90	96.8	Logit	89.9	90	89.9	97.1	RF	88.6	88	88	95.6
Dummy	0	0	0	50	Dummy	0	0	0	0	Dummy	0	0	0	50

the exercises. We also found in each FS method, the models had higher scores compared to the baseline dummy classifier (Table 2).

While exploring more the performance of the best classifier regardless of FS method, we found that the MLP model identified hand raised, wrist pronation, and supination 100% accurately (Table 3). However, in wrist flexion, though the predicted class was 100% accurate, it correctly identified 90% of exercises among 10 flexion exercises of 10 participants (precision = 100%, sensitivity = 90%, support = 10).

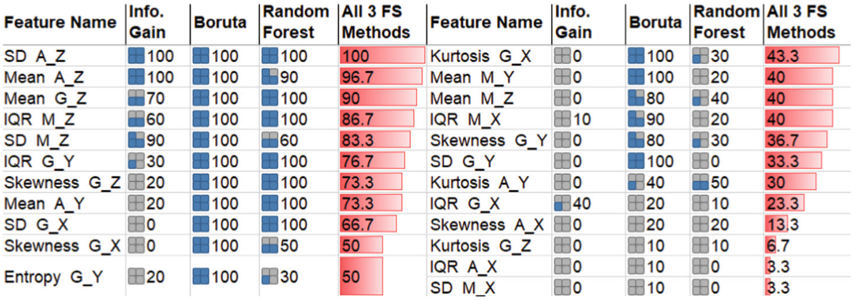
Table 3. Best model's (MLP based on 16 features selected by RF) prediction for each exercise.

Exercise	Precision	Sensitivity	F1	Support	Exercise	Precision	Sensitivity	F1	Support
Hand raised	100	100	100	10	Wrist pronation	100	100	100	10
Wrist flexion	100	90	95	10	Wrist supination	100	100	100	10
Wrist extension	91	100	95	10					

5.2 Feature Importance

We found 29 features (Fig. 7) that were used in the top-5 classifiers on the basis of the best set of features of each FS method (Table 2). Among them, 14 features (48.28%) were based on the gyroscope sensed data, which reflects that this sensor's features are more important for identifying the exercises (Fig. 7).

In the Boruta and RF FS methods, we found the stability of 6 features such as the mean and skewness of the gyroscope sensed data in the Z-axis, which appeared in all iterations of the LOOCV (Fig. 7). This may explain the fact of having relatively identical performance in RF and Boruta selected feature-based ML models. For example, the best model based on the RF selected features from 11 to 14 and also from 20 to 25, had identical performance (Fig. 6(b)). Also, at depths 4, 5, and 6 of the base estimator of Boruta FS, the performance was almost identical (Fig. 5(b)).



Feature Name	Info. Gain	Boruta	Random Forest	All 3 FS Methods	Feature Name	Info. Gain	Boruta	Random Forest	All 3 FS Methods
SD A_Z	100	100	100	100	Kurtosis G_X	0	100	30	43.3
Mean A_Z	100	100	90	96.7	Mean M_Y	0	100	20	40
Mean G_Z	70	100	100	90	Mean M_Z	0	80	40	40
IQR M_Z	60	100	100	86.7	IQR M_X	10	90	20	40
SD M_Z	90	100	60	83.3	Skewness G_Y	0	80	30	36.7
IQR G_Y	30	100	100	76.7	SD G_Y	0	100	0	33.3
Skewness G_Z	20	100	100	73.3	Kurtosis A_Y	0	40	50	30
Mean A_Y	20	100	100	73.3	IQR G_X	40	20	10	23.3
SD G_X	0	100	100	66.7	Skewness A_X	0	20	20	13.3
Skewness G_X	0	100	50	50	Kurtosis G_Z	0	10	10	6.7
Entropy G_Y	20	100	30	50	IQR A_X	0	10	0	3.3
					SD M_X	0	10	0	3.3

Fig. 7. Features used to develop the top-5 classifiers. Here, each value presents each feature's percentage of appearance in all 10 iterations of LOOCV. A: Accelerometer, M: Magnetometer, G: Gyroscope. X, Y, and Z denote the axes.

6 Discussion

We presented a low-cost system (~$16) to support UL rehabilitation in resource-constrained settings. Based on our system's sensed data, we developed ML models which can identify the 5 exercises with a ROC-AUC score of over 95%. This extends the existing systems, particularly the low-cost systems focusing on a few other exercises [1] where there is no automated process for accurate identification of exercises [1, 11]. Therefore, our system could identify and inform patients and caregivers whether the particular exercise is conducted precisely. This could facilitate home-based rehabilitation and support physiotherapists in remote monitoring, especially when there are inadequate rehabilitation facilities [16].

We found MLP as the best-performing model where its predicted exercise was accurate in 98.2% of cases. A plausible reason for the higher performance of MLP can be due to the neural networks' ability to capture complex patterns. But recent systematic reviews in medical informatics found researchers' preference for tree-based ML algorithms [13]. Though we developed models based on 8 tree-based algorithms, in the top-5 classifiers of each FS method, we found a single tree-based model. However, there were 3 models based on conventional ML algorithms such as the SVC and Logit where evaluation metrics' scores were over 90%. Conventional ML algorithms have fewer parameters that do not get overfitted easily. Also, considering the smaller sample size, we used nested cross-validation to build the models, which are found to prevent overfitting and show unbiased performance [22]. Hence, our findings suggest incorporating conventional ML algorithms along with complex algorithms while developing models to identify exercises for UL rehabilitation.

7 Limitations

The main limitations of our study are the low number of participants, especially with impairments in the upper extremities or undergoing physical rehabilitation. As the aim of this study was to evaluate the feasibility of our proof of concept system, future work should focus on applying more robust evaluation methods.

8 Conclusion

We presented an affordable system that was designed by integrating physiotherapists' perspectives. We presented the applicability of our system in accurately identifying 5 exercises with 10 participants to show its feasibility. Our system can play a role in home-based UL rehabilitation in low-resource settings such as Bangladesh.

Acknowledgment. We thank Dr. Catt Turney for input in the physiotherapist research, and all participants for their time. This study was supported by seed funding from Cardiff University's GCRF QR Funding from the Higher Education Funding Council for Wales.

References

1. Anowar, J., Ali, A.A., Amin, M.A.: A low-cost wearable rehabilitation device. In: Proceedings of the 2020 12th ICCAE. ACM (2020)
2. Bangladesh Bureau of Statistics. Household Income and Expenditure Survey 2016
3. Chakraborty, P.K., Islam, M.J., Hossain, M.S., Barua, S.K., Rahman, S.: Profile of patients receiving stroke rehabilitation in A tertiary care Hospital. Chattagram Maa-O-Shishu Hosp. Med. Coll. j. **17**, 9–12 (2018). https://doi.org/10.3329/cmoshmcj.v17i1.39435
4. D'Auria, D., Persia, F., Siciliano, B.: Human-computer interaction in healthcare: how to support patients during their wrist rehabilitation. In: 2016 IEEE Tenth ICSC. IEEE (2016)
5. Dutta, D., Sen, S., Aruchamy, S., Mandal, S.: Prevalence of post-stroke upper extremity paresis in developing countries and significance of m-Health for rehabilitation after stroke - a review. Smart Health **23**, 100264 (2022). https://doi.org/10.1016/j.smhl.2022.100264
6. boruta_py. https://github.com/scikit-learn-contrib/boruta_py. Accessed 03 Aug 2022
7. Tran, P., Jeong, S., Wolf, S.L., Desai, J.P.: Patient-specific, voice-controlled, robotic FLEXotendon glove-II system for spinal cord injury. IEEE Robot. Autom. Lett. (2020)
8. EVAR - Noise variance estimation. https://www.biomecardio.com/matlab/evar_doc.html. Accessed 05 Aug 2022
9. Grossman, R., Salas, E.: The transfer of training: what really matters: the transfer of training. Int. J. Train. Dev. **15**, 103–120 (2011)
10. Hubbard, I.J., Parsons, M.W., Neilson, C., Carey, L.M.: Task-specific training: evidence for and translation to clinical practice: task-specific training in clinical practice. Occup. Ther. Int. **16**, 175–189 (2009). https://doi.org/10.1002/oti.275
11. Hughes, C.M.L., et al.: Development of a post-stroke upper limb rehabilitation wearable sensor for use in sub-Saharan Africa: a pilot validation study. Front. Bioeng. Biotechnol. **7**, 322 (2019)
12. Ietswaart, M., et al.: Mental practice with motor imagery in stroke recovery: randomized controlled trial of efficacy. Brain **134**, 1373–1386 (2011)

13. Brnabic, A., Hess, L.M.: Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Med. Inform. Decis. Mak. **21**, 54 (2021). https://doi.org/10.1186/s12911-021-01403-2
14. Sunderland, A., Tuke, A.: Neuroplasticity, learning and recovery after stroke: a critical evaluation of constraint-induced therapy. Neuropsychol. Rehabil. **15**, 81–96 (2005)
15. Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A. (eds.): Feature Extraction: Foundations and Applications. Springer, Heidelberg (2006)
16. Uddin, T., Islam, M.T., Rathore, F.A., O'Connell, C.: Disability and rehabilitation medicine in Bangladesh: current scenario and future perspectives. J. Int. Soc. Phys. Rehabil. Med. (2019)
17. Low, K.S., Lee, G.X., Taher, T.: A wearable wireless sensor network for human limbs monitoring. In: 2009 IEEE I2MTC. IEEE (2009)
18. Hsieh, K.L., Sosnoff, J.J.: Smartphone accelerometry to assess postural control in individuals with multiple sclerosis. Gait Posture **84**, 114–119 (2021)
19. Maciejasz, P., Eschweiler, J., Gerlach-Hahn, K., Jansen-Troy, A., Leonhardt, S.: A survey on robotic devices for upper limb rehabilitation. J. Neuroeng. Rehabil. **11**, 3 (2014)
20. Mamin, F.A., Islam, M.S., Rumana, F.S., Faruqui, F.: Profile of stroke patients treated at a rehabilitation centre in Bangladesh. BMC Res. Notes **10**, 520 (2017)
21. Mawson, S., Nasr, N., Parker, J., Davies, R., Zheng, H., Mountain, G.: A personalized self-management rehabilitation system with an intelligent shoe for stroke survivors: a realist evaluation. JMIR Rehabil. Assist. Technol. **3**, e1 (2016)
22. Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J.: Machine learning algorithm validation with a limited sample size. PLoS ONE **14**, e0224365 (2019)
23. Kursa, M.B., Rudnicki, W.R.: Feature selection with the Boruta package. J. Stat. Softw. (2010). https://www.jstatsoft.org/article/view/v036i11
24. Electrical Stimulation - Its role in upper limb recovery post-stroke. https://www.physio-pedia.com/index.php?title=Electrical_Stimulation_-_Its_role_in_upper_limb_recovery_post-stroke&oldid=216559. Accessed 01 Aug 2022
25. Maceira-Elvira, P., Popa, T., Schmid, A.-C., Hummel, F.C.: Wearable technology in stroke rehabilitation: towards improved diagnosis and treatment of upper-limb motor impairment. J. Neuroeng. Rehabil. **16**, 142 (2019). https://doi.org/10.1186/s12984-019-0612-y

Computer Vision

Multiclass Semantic Segmentation of Mediterranean Food Images

Fotios S. Konstantakopoulos[1], Eleni I. Georga[1], and Dimitrios I. Fotiadis[1,2](✉)

[1] Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, Ioannina, Greece
`{fotkonstan,egeorga,fotiadis}@uoi.gr`
[2] Biomedical Research Institute, FORTH, Ioannina, Greece

Abstract. With the continuous increase of artificial intelligence applications in modern life, the segmentation of images is one of the fundamental tasks in computer vision. Image segmentation is the key for many applications and is backed by a large amount of research, including medical image analysis, healthcare services and autonomous vehicles. In this study we present a semantic segmentation model for food images, suitable for healthcare systems and applications as major part of the dietary monitoring pipeline, trained on an annotation dataset of Mediterranean cuisine food images. To segment the images, we use for feature extraction the ResNet-101 CNN model pre-trained on the ImageNet LSVRC-2012 dataset as a backbone network and the Pyramid Scene Parsing Network - PSPNet architecture for food image segmentation. For the evaluation metric we use the Intersection over Union, where the proposed model achieves a meanIoU score 0.758 in 50 classes of the Mediterranean Greek Food image dataset and 0.933 IoU score in food/non-food segmentation. To evaluate the proposed segmentation model, we train and evaluate a U-Net segmentation model on the same dataset, which achieves meanIoU 0.654 and IoU score 0.901 in multiclass and food/non-food segmentation, respectively.

Keywords: Computer Vision · Image Segmentation · Semantic Segmentation · Deep Learning · Food Image Dataset · Dietary Assessment Systems

1 Introduction

In healthcare systems, image segmentation is used to segment biomedical images into different regions to assist physicians in diagnosing diseases [1]. Also, image segmentation can be used to dietary assessment applications, as part of the nutritional composition system, to assist individuals to follow a healthy diet, preventing chronic diseases such as obesity, diabetes, cardiovascular diseases (CVDs) and cancer [2]. Every year, millions of people are died from chronic diseases. For example, in 2021, diabetes was responsible for 6.7 million deaths worldwide [3]. A common factor that can affect the treatment of the above diseases is the management of the daily diet. Healthy habits are essential for the management of these diseases and, in some cases, changing the daily diet may be enough to control the disease.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 49–59, 2023.
https://doi.org/10.1007/978-3-031-34586-9_4

The image dataset is the key to creating a high-accurate model for a food image segmentation system. However, image datasets for deep learning segmentation models are hard to collect, because a lot of professional expertise is needed to label them. Moreover, the need for highly performance deep learning models requires the collection of large numbers of images. In dietary assessment systems there are a few datasets suitable for the training and evaluation of deep learning segmentation models. A food image dataset can be characterized by the total number of images they include, the number of food classes, the source of the food images, the type of cuisine and by the task they can be used (e.g., food segmentation, food classification or food volume estimation task). For example; Food524DB [4] represents a generic type of cuisine and consists of 247,636 food images with 524 food classes acquired from previous datasets; Vireo Food-172 [5] represents the Japanese cuisine and consists of 110,241 food images with 172 food classes downloaded from the web; while Food201-segmented [6] can be used for food image segmentation tasks.

The food image segmentation task plays an important role in AI applications for the daily management of nutrition [7]. These systems are divided in two main categories: (i) traditional machine learning segmentation approaches with handcrafted feature extraction [8], and (ii) deep learning segmentation approaches with automatic feature extraction [9]. Traditional machine learning approaches, use feature extraction algorithms, such as Gabor features and Speed-up robust features (SURF), to find and extract the features of the food image and then, the features are fed to a classifier, such as random forests, to segment the food [10]. In [11], an interactive food image segmentation algorithm has been proposed, where food parts are extracted based on user's inputs in the first step and then, a boundary detection and filling and the Gappy principal component analysis methods are applied to restore the missing information.

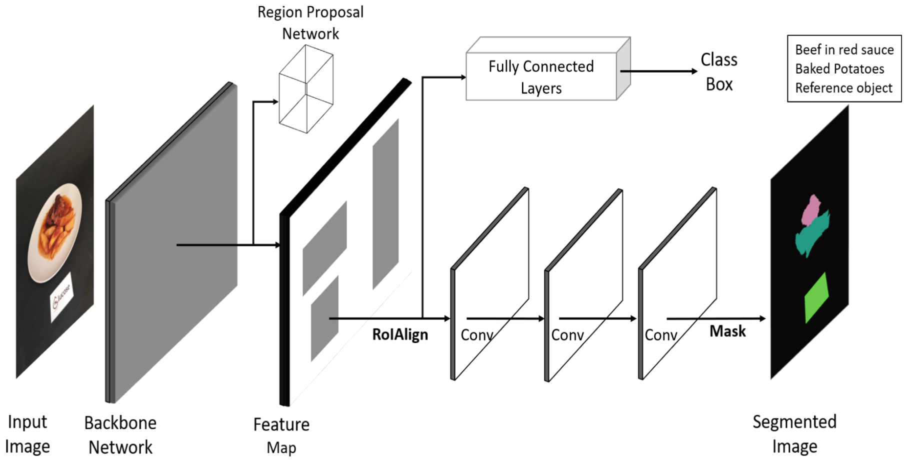


Fig. 1. An instance segmentation model for food items pixel classification.

Today, Convolutional Neural Networks (CNNs) a class of deep neural networks (DNN), are the state of the art methodology in food image segmentation systems. CNN models are extremely accurate for computer vision tasks and surpass the traditional machine learning models in image segmentation Intersection over Union (IoU) metric. The deep learning techniques that are used for image segmentation are divided into semantic and instance segmentation techniques (Fig. 1). Semantic segmentation models provide segment maps as outputs that correspond only to the inputs they are fed, while instance segmentation models detect and delineate each distinct object of interest that appears in the image. In food image segmentation, semantic segmentation techniques are mainly used, which aim either to segment the food from the background, or to segment the different types of food contained in the image. For example, in [12] they proposed a semantic segmentation deep learning model which achieved 0.931 IoU score in food/no-food segmentation task using a DeepLab-V2 model in the UNIMIB-2016 [13] food image dataset, while in [9] they achieved 0.439 meanIoU for the segmentation of 103 food labels, by combining the proposed Recipe Learning Module (ReLe) and the Segmentation Transformer (SeTR) [14].

In this study, we propose a semantic segmentation network to segment images containing Mediterranean foods. Using a new food image annotated dataset, we present the architecture of the proposed segmentation model and its training pipeline. Our network is suitable for detecting specific food items as well as for separating food from the background using a semantic segmentation model. Although the segmentation step is not necessary in several dietary assessment systems, we observe that the studies using the segmentation step, result in better performance [15]. Relative to similar approaches, the innovation of this study is that it proposes a state of the art pre-trained DCNN model for food image segmentation using a novel annotated dataset of Mediterranean cuisine food images. While most related approaches focus on either handcrafted feature extraction or using existing annotated datasets [15], we propose a deep learning model for feature extraction and a new annotated food image dataset, which can be used in classification and volume estimation stages in dietary assessment systems. The proposed model can be part of dietary assessment systems and applications, by improving the accuracy of image classification and food volume estimation stages. In addition, it is suitable for healthcare systems that monitor patient malnutrition in hospitals, offering the ability to track the different food items served with the tray to the patient. Finally, to prove the dominance of the proposed segmentation model, we compare the performance of an additional segmentation model using a different architecture.

2 Methods

2.1 Food Image Dataset

In the present study, for the training and the evaluation of the segmentation model, we use two image datasets: (i) the ImageNet LSVRC-2012, and (ii) the MedGRFood[1] [16]. ImageNet is a large image dataset, that contains 1,431,167 images belonging to 1,000 object classes, such as bus, dog, puzzle etc. The MedGRFood is a new food image dataset, which contains 51,840 Mediterranean food images belonging to 160 classes appropriate for classification tasks and an additional 20,000 Mediterranean food images belonging to 190 classes appropriate for volume estimation tasks. All the images have been collected from the web and under a controlled environment along with their weight. For the proposed segmentation model, we annotated 5,000 food images of 50 classes from the MedGRFood dataset, with respect to the food category, the exact food name, the cuisine and the weight of food. Figure 2 shows images from the datasets ImageNet and MedGRFood.

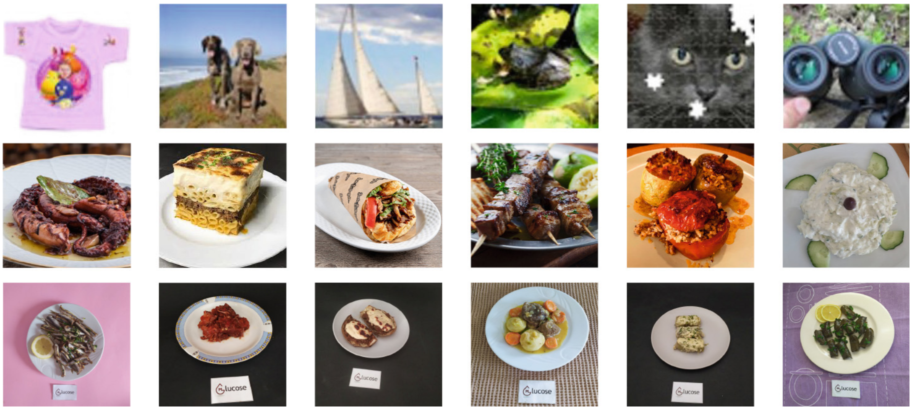


Fig. 2. Images from ImageNet LSVR and MedGRFood datasets. The first row shows images from the ImageNet dataset and the others rows show images from the MedGRFood dataset.

2.2 Segmentation Network Architecture

Knowing that most semantic segmentation models contain two main parts (i) an encoder and (ii) a decoder, for feature map extraction and pixel prediction, respectively, we choose the Pyramid Scene Parsing Network (PSPNet) [17] architecture for food image segmentation. Specifically, the proposed PSPNet encoder contains the ResNet-101 [18] model as backbone network with dilated convolutions along with the pyramid pooling module for feature extraction. ResNet-101 is a 101-layer deep CNN that democratizes the concepts of residual learning and skip the connections between some of the blocks.

---

[1] MedGRFood dataset is available for research purposes via the website: http://glucoseml.gr/.

We use a pretrained version of ResNet-101, trained on the ImageNet dataset. Knowing that the early layers on CNNs extract and learn general features (such as edges and simple textures) while the later layers extract and learn detailed or high-level features (such as more complex textures and patterns), we take advantage of the ImageNet dataset by transferring knowledge to our own task. In PSPNet architecture, the last layers of the backbone network replace the convolutional layers with dilated convolutional layers, which help the receptive field to grow. The dilated convolution layers are placed in the last two blocks of the backbone network with dilation values two and four, respectively, so that the features obtained at the end of the backbone contain richer features. The dilation value determines the sparsity when performing the convolution. The pyramid pooling module is the main part of the PSPNet architecture, which acts as an efficient global contextual prior, helping the model to classify the pixels based on the global information present in the image. Using a multi-level pyramid with four different scales (1 × 1, 2 × 2, 3 × 3 and 6 × 6), the pooling kernels cover different size portions of the image. After each pyramid level, we used a 1 × 1 convolutional layer to reduce the dimension of context to maintain the weight of global feature. Then, the upsampled maps are concatenated with the original feature map to pass to the decoder. The decoder takes the features of the encoder and turns them into predictions, by passing them into its layers. Finally, we use a convolutional layer followed by an 8x bilinear upasampling, as decoder for our segmentation network to recover the original size. The architecture of the proposed classification model is shown in Fig. 3.

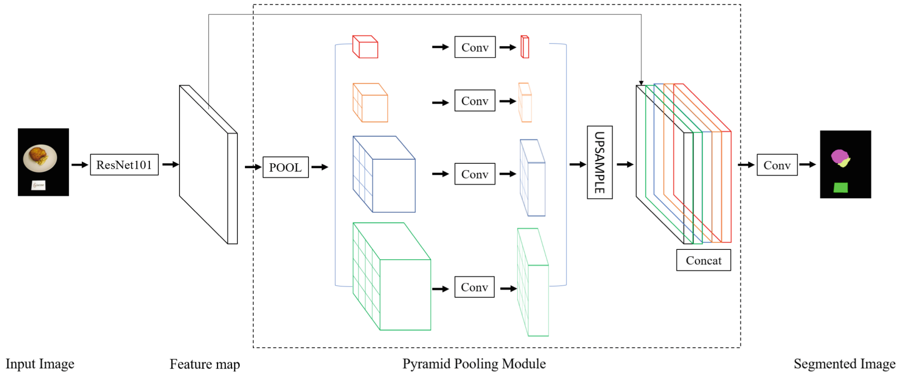


Fig. 3. The architecture of the proposed semantic segmentation model

2.3 Training and Testing

For the training phase, all images are resized to 240 × 240 × 3 resolution and the total number of the models' trainable parameters are 4,052,219. The MedGRFood dataset is partitioned into the training and validation set, using a ratio of 90:10 (90% is used for model training and 10% is used for model validation). In total, we used 4,500 food images and their masks for training, and 677 images and their masks for the validation set. Moreover, we chose a scaled learning rate with initial value 0.0001 and final value 0.0000001. The learning rate decreases by a factor 0.9 when the validation loss stops improving for three epochs. The model is trained for 50 epochs using the Adam optimizer [19], with an early stopping function if the validation accuracy stops improving for five epochs.

In several studies [20], pixel accuracy is chosen as the evaluation metric for the segmentation task. This metric can sometimes provide misleading results, when there are classes in the image with few pixel representations, as the measure will be biased in reporting how well you recognize the negative case. Here, we used the mean Intersection over Union score (meanIoU) to compute the accuracy of the proposed segmentation model. The IoU measures the similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets (Eq. 1). Then, to calculate the meanIoU of the 50 food classes we used Eq. 2:

$$IoU = \frac{Y_{true} \cap Y_{pred}}{Y_{true} \cup Y_{pred}}, \tag{1}$$

$$meanIoU = \frac{1}{50} \sum\nolimits_{i=1}^{50} IoU_i, \tag{2}$$

where Y_{true} is the ground truth of the food image and Y_{pred} is the prediction mask.

We used the IoU loss function for our segmentation problem. The Eq. 3 shows the computation of IoU_{loss}:

$$IoU_{loss} = 1 - IoU. \tag{3}$$

2.4 Implementation

We used the python programming language to implement the semantic segmentation model in the Anaconda environment. Knowing the increased computing power requirements of CNN models, we also used the cuda toolkit, the cudnn and tensorflow libraries, for model training and validation of classification subsystem, through the Nvidia GeForce RTX 3080 graphic processing unit. Also, we used the opencv and the segmentation models libraries for the implementation of the proposed food segmentation models.

3 Results

To evaluate the proposed food semantic segmentation model, we further constructed and trained an additional segmentation model using the U-net architecture [21]. We trained the U-net model with exactly the same parametrization that we applied to the proposed model with the PSPNet architecture. Moreover, we built and trained two additional binary segmentation models for food and non-food segmentation. At these models we aimed to segment food regions from the background. Table 1 presents the segmentation results of the four models. We observe that the proposed model achieves a higher meanIoU score from the U-net model. The PSPNet architecture considers the global context of the image to predict the local level prediction and, therefore, gives better performance on the MedGRFood dataset. In addition, the difference in the total number of generated parameters between the two models is very large, which requires more training time for the U-net model.

Table 1. Segmentation Results Between the Four Models.

Model	MeanIoU	Loss	Training time (ms/step)	Number of parameters ($x10^6$)
Multiclass PSPNet	0.758	0.242	320	4.052
Multiclass U-net	0.654	0.346	370	51.512
Binary PSPNet	0.933	0.076	140	4.052
Binary U-net	0.901	0.099	201	51.512

In Fig. 4 we present the multiclass and binary segmentation results of the proposed model. We observe that the predicted mask is very close to the real mask of the test image in multiclass segmentation. We also notice that there is no wrong result in the class estimation, despite the fact that the number of 50 classes is quite large. We see that there is a slight deviation in the food mask prediction from its ground truth. Regarding the food segmentation from the image background, i.e., the binary segmentation, the extracted food mask is almost identical to its actual mask. This is a very crucial step in dietary assessment systems, because having the mask of the food we can use a classification model to predict its class very accurately. In Fig. 5 we present the multiclass and binary segmentation results of the U-net model. In multiclass semantic segmentation,

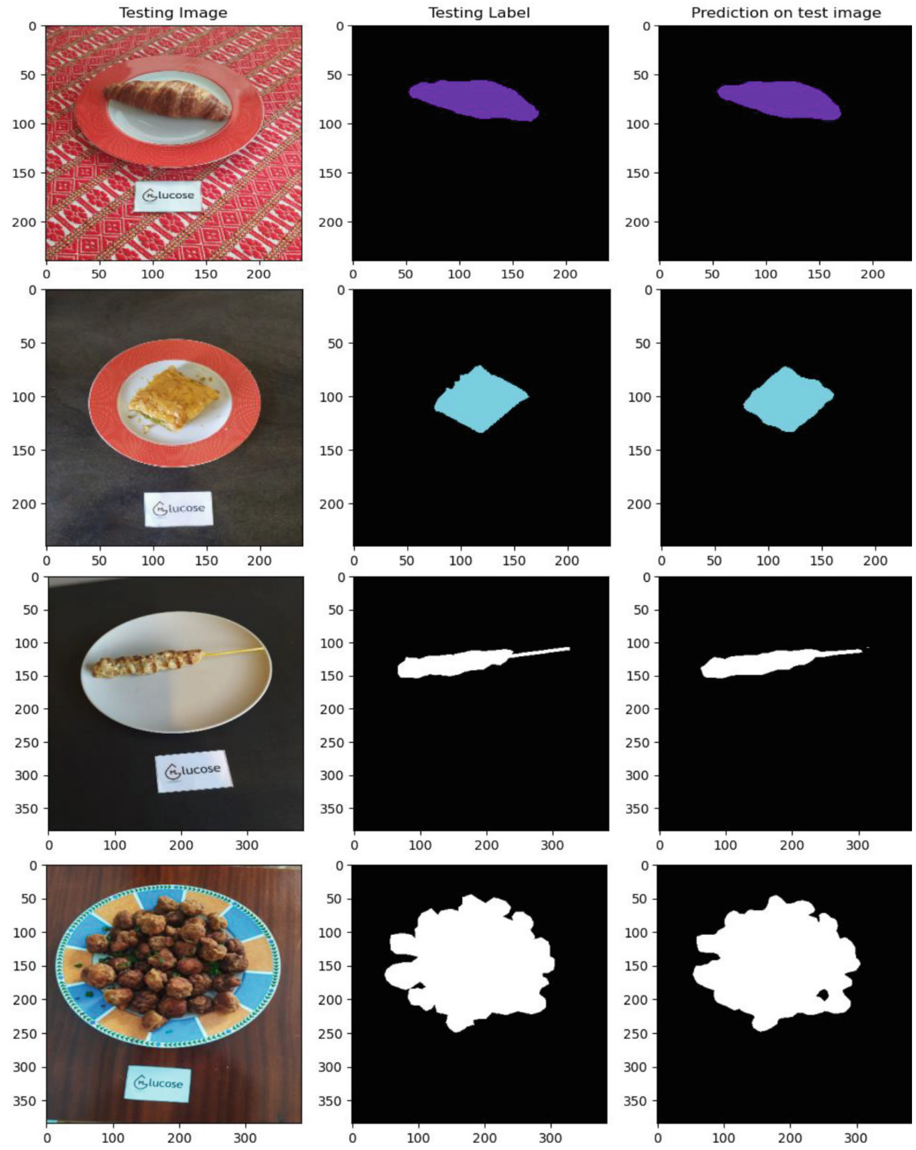


Fig. 4. Multiclass and binary semantic segmentation results of the proposed model.

we observe that the predicted mask has differences from the food ground truth. Moreover, we see that the U-net model incorrectly recognizes the food class in the second row. Comparing the results of the two semantic segmentation architectures, we can say that U-net does not perform as well as PSPNet, as it is not able to capture the context of the whole image. In predicted masks with differences with their ground truths, the application of morphological operations, such as erosion and dilation, could lead to the improvement of the results.

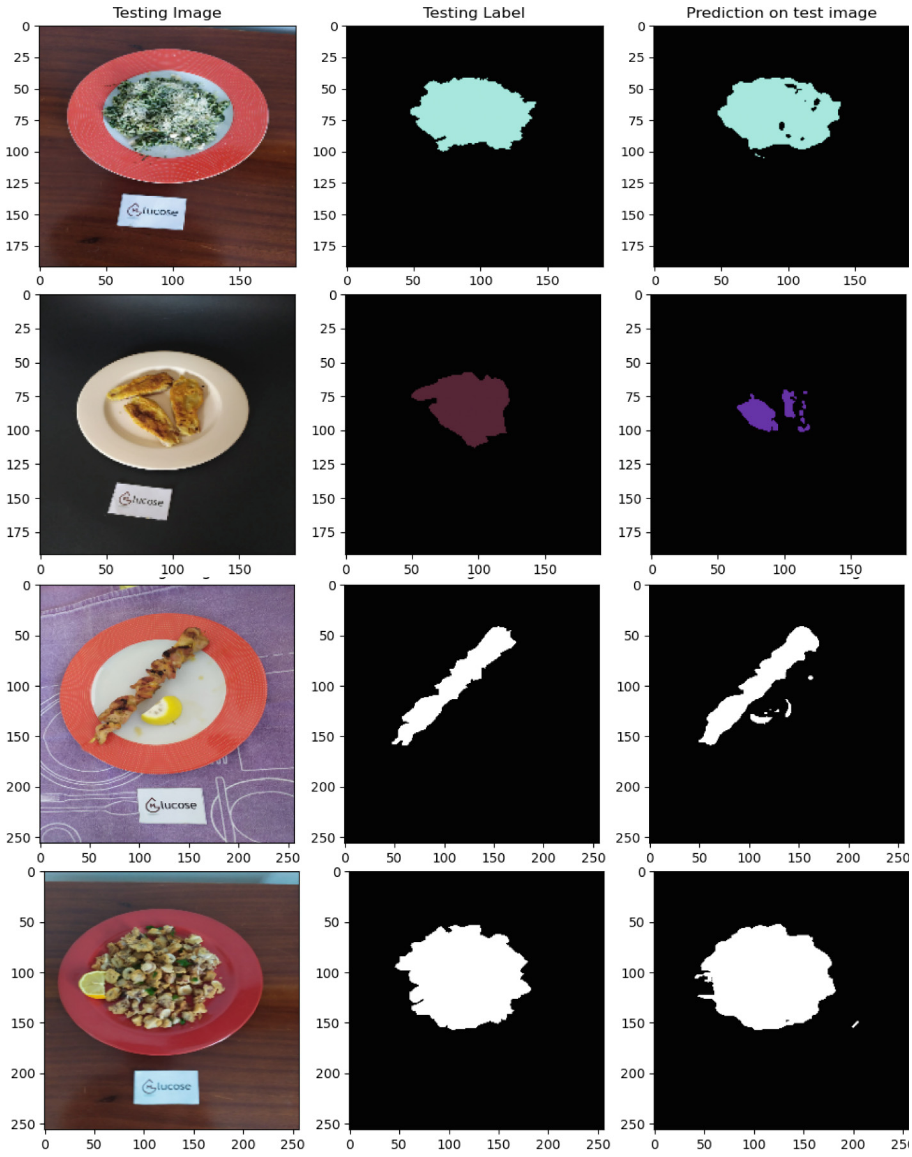


Fig. 5. Multiclass and binary semantic segmentation results of the U-net model.

4 Discussion and Conclusions

In food image databases, the use of deep learning techniques for food segmentation tends to create annotated databases with the largest possible number of images for each food class. However, the existing annotated databases are few and limited to the number of food classes they contain. Thus, there is a necessity to create a generic annotated food image database which covers as many food categories as possible and represents the types of food from all cuisines. The collection of food images and the creation of food image databases is an easier task nowadays, due to the habit of capturing and posting images on social media. However, creating an annotated database of food images using their weight in addition to the type of food, remains a challenging task and will help build better and more accurate models for the segmentation and volume estimation steps in dietary assessment systems.

In automated food segmentation, the use of deep learning techniques has resulted in better performance compared to image processing techniques. Semantic segmentation and instance segmentation are techniques that have been used on a small scale in food image segmentation and could further improve the segmentation performance of dietary assessment systems. This presupposes the use of annotated food image databases, as it is a prerequisite to build segmentation models based on deep learning. In recent studies [22], the step of food image segmentation is omitted and in some others the performance of this step is not reported. In other studies, although the performance of the methods used to segment food images is high and improves the classification accuracy, there are still open issues related to cases where there are mixed foods. In these cases, the use of state of the art segmentation techniques, such as semantic and instance segmentation, can be used to improve the performance of this step and improve the efficiency to the classification step.

In this study, we presented a semantic segmentation model for multiclass and binary segmentation, using the pre-trained ResNet-101 as backbone network to the PSPNet architecture, applying the transfer learning technique from the ImageNet dataset. Comparing our results with related studies, we notice that the meanIoU score for multiclass segmentation has an excellent value, while the IoU score for food/non-food segmentation is one of the best results in the related literature [15]. To demonstrate the superiority of the proposed methodology, we built and trained an additional segmentation model based on the U-net architecture. The proposed model performs better and provides more accurate food segments in both multiclass and binary segmentation. This is due to the PSPNet ability to render the context of the whole image and to locate the objects of interest with higher accuracy. Open issues of this study are: (i) the ability of the segmentation model to separate complex foods, (ii) the segmentation of dishes containing two or more food items, (iii) the segmentation of dishes with overlaps between food items and, (iv) to calculate a good accuracy score of semantic segmentation models on an image with two or more food classes.

Acknowledgments. This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness. Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EDK-03185).

References

1. Minaee, S., Boykov, Y.Y., Porikli, F., Plaza, A.J., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: a survey. J. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 3523–3542 (2021)

2. Farràs, M., et al.: Beneficial effects of olive oil and Mediterranean diet on cancer physio-pathology and incidence. Semin. Cancer Biol. **73**, 178–195 (2021)
3. https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html. Accessed 9 Dec 2021
4. Ciocca, G., Napoletano, P., Schettini, R.: Learning CNN-based features for retrieval of food images. In: Battiato, S., Farinella, G.M., Leo, M., Gallo, G. (eds.) ICIAP 2017. LNCS, vol. 10590, pp. 426–434. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70742-6_41
5. Chen, J., Ngo, C.-W.: Deep-based ingredient recognition for cooking recipe retrieval, pp. 32–41 (2016)
6. Meyers, A., et al.: Im2Calories: towards an automated mobile vision food diary, pp. 1233–1241 (2015)
7. Konstantakopoulos, F.S., et al.: GlucoseML mobile application for automated dietary assessment of mediterranean food, pp. 1432–1435. IEEE (2022)
8. Fang, S., Liu, C., Tahboub, K., Zhu, F., Delp, E.J., Boushey, C.J.: cTADA: the design of a crowdsourcing tool for online food image identification and segmentation, pp. 25–28. IEEE (2018)
9. Wu, X., Fu, X., Liu, Y., Lim, E.-P., Hoi, S.C., Sun, Q.: A large-scale benchmark for food image segmentation, pp. 506–515 (2021)
10. Pouladzadeh, P., Shirmohammadi, S., Bakirov, A., Bulut, A., Yassine, A.: Cloud-based SVM for food categorization. Multimed. Tools Appl. **74**(14), 5243–5260 (2015)
11. Inunganbi, S., Seal, A., Khanna, P.: Classification of food images through interactive image segmentation. In: Nguyen, N.T., Hoang, D.H., Hong, T.-P., Pham, H., Trawiński, B. (eds.) ACIIDS 2018. LNCS (LNAI), vol. 10752, pp. 519–528. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75420-8_49
12. Aslan, S., Ciocca, G., Schettini, R.: Semantic food segmentation for automatic dietary monitoring, pp. 1–6. IEEE (2018)
13. Ciocca, G., Napoletano, P., Schettini, R.: Food recognition: a new dataset, experiments, and results. IEEE J. Biomed. Health Inform. **21**(3), 588–598 (2016)
14. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, pp. 6881–6890 (2021)
15. Wang, W., et al.: A review on vision-based analysis for automatic dietary assessment. Trends Food Sci. (2022)
16. Konstantakopoulos, F., Georga, E.I., Fotiadis, D.I.: 3D reconstruction and volume estimation of food using stereo vision techniques, pp. 1–4. IEEE (2021)
17. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network, pp. 2881–2890 (2017)
18. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: revisiting the ResNet model for visual recognition. J. Pattern Recogn. **90**, 119–133 (2019)
19. Zhang, Z.: Improved Adam optimizer for deep neural networks, pp. 1–2. IEEE (2018)
20. Subhi, M.A., Ali, S.H., Mohammed, M.A.: Vision-based approaches for automatic food recognition and dietary assessment: a survey. IEEE Access **7**, 35370–35381 (2019). https://doi.org/10.1109/ACCESS.2019.2904519
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
22. Yunus, R., et al.: A framework to estimate the nutritional value of food in real time using deep learning techniques. J. IEEE Access **7**, 2643–2652 (2018)

Comparative Study of Machine Learning Methods on Spectroscopy Images for Blood Glucose Estimation

Tahsin Kazi[1], Kiran Ponakaladinne[1], Maria Valero[1](✉), Liang Zhao[1], Hossain Shahriar[1], and Katherine H. Ingram[2]

[1] Department of Information Technology, Kennesaw State University, Kennesaw, GA 30060, USA
`mvalero2@kennesaw.edu`
[2] Department of Exercise Science and Sport Management, Kennesaw State University, Kennesaw, GA 30060, USA

Abstract. Diabetes and metabolic diseases are considered a silent epidemic in the United States. Monitoring blood glucose, the lead indicator of these diseases, involves either a cumbersome process of extracting blood several times per day or implanting needles under the skin. However, new technologies have emerged for non-invasive blood glucose monitoring, including light absorption and spectroscopy methods. In this paper, we performed a comparative study of diverse Machine Learning (ML) methods on spectroscopy images to estimate blood glucose concentration. We used a database of fingertip images from 45 human subjects and trained several ML methods based on image tensors, color intensity, and statistical image information. We determined that for spectroscopy images, AdaBoost trained with KNeigbors is the best model to estimate blood glucose with a percentage of 90.78% of results in zone "A" (accurate) and 9.22% in zone "B" (clinically acceptable) according to the Clarke Error Grid metric.

Keywords: Non-invasive monitoring · spectroscopy · machine learning · blood glucose concentration

1 Introduction

The USA is facing an epidemic of diabetes [26], with more than 11.33% of the population affected by diabetes alone [7] and another 30% by metabolic syndrome [16]. The lead indicator of these diseases is the blood glucose (BG) concentration. Measuring blood glucose typically involves either painful blood extraction multiple times per day or implanting needles under the skin for continuous monitoring.

Supported by College of Computing and Software Engineering and Office of Research at Kennesaw State University.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 60–74, 2023.
https://doi.org/10.1007/978-3-031-34586-9_5

Non-invasive estimation of blood glucose levels has emerged as an exciting alternative for the monitoring and management of metabolic diseases [33]. Among those, technologies that use optical approaches are both practical and inexpensive [2,13,23,31]. Optical methods function by directing a light beam through human tissue, and the energy absorption, reflection, or scattering is used to estimate blood glucose concentration [20]. Optical methods are portable and can be easily applied to fingers and other extremities such as earlobes. Several optical methods for detecting blood glucose have been developed, though most of these have limitations that restrict their utility. They include: (1) fluorescence spectroscopy [19], which may result in harmful exposure to fluorophore [12]; (2) Raman spectroscopy [10] criticized for its lengthy spectral acquisition time and poor signal-to-noise ratio [32]; (3) photoacoustic spectroscopy [18], which introduces noise from its sensitivity to environmental factors [32]; (4) optical coherence tomography [11], which is overly sensitive to skin temperature; and (5) occlusion spectroscopy [4], known to result in signal drift [27]. An alternative optical method (6) near-infrared absorption spectroscopy, avoids these limitations and is both more practical and cost-efficient than those described above [2,13,15,21,23,31]. In addition, near-infrared absorption spectroscopy is fundamentally simple to use in the creation of a powerful sensor prototype. Just a laser light and camera are needed.

In our previous work [31], we designed and tested a non-invasive sensor prototype for estimating blood glucose based on near-infrared absorption spectroscopy. A device composed of laser light, a raspberry Pi, and a camera was used to collect fingertip images, as shown in Fig. 1. These images were used to estimate blood glucose by applying a Machine Learning (ML) model, specifically a Convolutional Neural Network (CNN). The model used light absorption data from the images to approximate a function for estimating blood glucose concentration. However, the initial accuracy only reached 72% with the limited dataset and that particular CNN model. Therefore, more studies were needed to determine a suitable ML method that provides higher accuracy for blood glucose estimation.

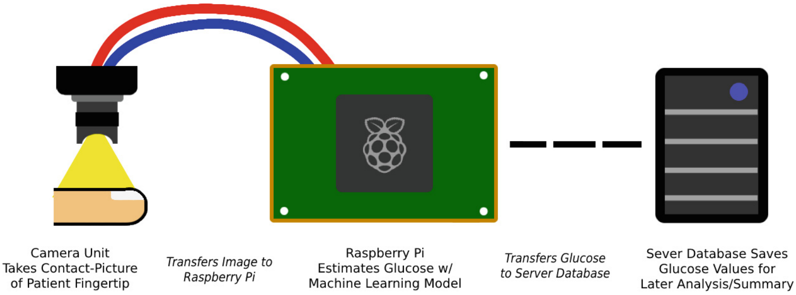


Fig. 1. Demonstration of the Blood Glucose Measuring System

This paper presents a comparative study of diverse ML methods trained with spectroscopy image data to identify the best model for estimating blood glucose. Metrics including mean absolute error (MAE), root mean square error (RMSE), and Clark error grids were used to determine accuracy. We further discuss the data preprocessing methods used to feed diverse ML models with the same dataset.

2 Background

In this section, we provide a background of the neural networks and linear statistical models applied to the spectroscopy image data obtained by our prototype [31], the CNN models VGG16 [30] and MobileNetV2 [28] were used to determine blood glucose concentration. On the other side, several linear models were employed, including Random Forest [9], Support Vector Machine [22], Bayesian Ridge [24], XGBoost [6], AdaBoost Ensemble [1], Histogram Gradient Boosting [5], Elastic Net [34], and KNeighbors. To apply these methods to the spectroscopy images, we used data transformation techniques to create new suitable databases for each method (Sect. 3). CNN models can only be used on tensor data because the algorithms are based on Linear Algebra that are suitable only for use with multi-dimensional matrices (tensors). Linear models are suitable for all scalar data and use a wide variety of statistical techniques to approximate the function of the data.

2.1 CNN Models

CNN models work by passing filters through images (represented as tensors) to extract features such as edges, shapes, and colors. These two-dimensional features are then flattened and mapped as scalar data which is then processed through normal neural network layers [3]. CNN models can use different types of filters through images of varying sizes, providing a wide range of applications. Since spectroscopy images were used, we applied CNN to determine blood glucose concentration using VGG16 and MobileNetV2.

VGG16 Neural Network: VGG-16 is a 16 layered deep CNN. A pretrained version of the network can be loaded which is trained on more than a million images from the ImageNet database [25]. The pretrained network can classify images into 1000 object categories. As a result, the network has learned rich feature representations for a wide range of images. However, the model was changed to output a single numeric value (blood glucose) instead of the 1000 categories it was trained on. The network has an image input size of 224-by-224, however the model's input size was changed to fit 160-by-120. VGG16 is well-suited for this project due to its ability to detect many different features and patterns as well as its performance when compared to other models [30]. An example of the VGG-16 architecture can be found in Fig. 2 (Taken from [29]).

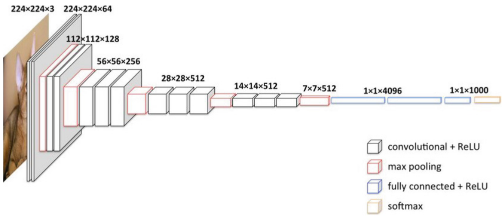


Fig. 2. Demonstration of VGG-16 Model Architecture, from [29].

MobileNetV2: MobileNetV2 is a mobile architecture that enhances the state-of-the-art performance of mobile models across various model sizes, tasks, and benchmarks. In contrast to conventional residual models, which use expanded representations for the input, the MobileNetV2 architecture is based on an inverted residual structure, where the input and output of the residual block are thin bottleneck layers. Although the architecture of this model is more complex than most other CNN models, it performs well considering its computational power. Therefore, it was possible to train MobileNetV2 normally instead of using a pre-trained model version. MobileNetV2 was chosen for this study because of its low computational power usage, fast training times, and high-performance [28].

2.2 Linear Statistical Models

Linear models are a staple of machine learning and statistical modeling due to the countless algorithms available for function approximation, decision making, regression, classification, clustering, and prediction. Like CNNs, linear models were also chosen due to their wide range of applications and their superior performance. They are significantly faster and less computationally intensive than neural networks, but they can provide similar or better results in many instances. Many of the linear models used in this study applied bagging, boosting, or ensemble learning techniques, which allow for higher performance, lower error, and more optimized training. We propose a mix of models using these techniques to determine the most effective for estimating blood glucose.

Random Forest Regressor (RFR). Random Forest is a supervised learning algorithm built on Decision Trees and the Ensemble Learning Approach [35]. Decision Trees are tree-diagrams of statistical decisions that lead users to a specific outcome, result, or prediction. Random Forest uses an optimized approach to ensemble learning called bagging (bootstrap-aggregating), which works like this: the model creates multiple decision trees that train on random segments of the training data, these trees are then used in unison to predict unknown values. Random Forest was chosen for this study for its novel combination of Decision Trees and bagging, and its high performance in many domains [9].

Support Vector Regressor (SVR). Support Vector Regression [22] works on the principle of the Support Vector Machine (SVM) [17]. This model is based on simple regression algorithms, to fit a line/curve/plane through the data to create an approximate function. In simple regression, the goal is to minimize the error rate while in SVR it is to fit the error inside a certain threshold. The flexibility of SVR allows to decide how much error is acceptable in the model, and it will find an appropriate line (or curve or plane) to fit the data accordingly. This technique was included for its ability to reduce overfitting and handle outliers in data. It is a well-performing and versatile model.

Bayesian Ridge Regressor (BRR). Ridge Regression is a classical regularization technique widely used in Statistics and ML [24]. Bayesian regression allows a natural mechanism to survive insufficient or poorly distributed data by generalizing the data, which significantly reduces overfitting and handles outliers. In addition, this model outputs with a probability distribution, which means that it outputs multiple predicted values and chooses the most likely value. This method was used in this study because it performs well regardless of data quality.

XGBoost Regressor (XGB). XGBoost uses gradient boosting, an ensemble learning using boosting. It trains multiple decision trees to create an ensemble learner and it relies on the intuition that the best possible next model, combined with previous models, minimizes the overall prediction error. Through combining multiple models training, the model achieves high performance, even in cases where insufficient data and outliers exist. Extreme Gradient Boosting (XGBoost) is an efficient open-source implementation of this gradient boosting algorithm [6]. The two benefits of using XGBoost are training speed and model performance, which is why it is chosen for this study.

Histogram Gradient Boosting Regressor (HGB). Histogram-based gradient boosting is an algorithm that uses the same gradient boosting as XGBoost, but instead of outputting a single value for blood glucose, it employs binning. Binning is a technique that converts continuous values into categories, similar to those used in classification scenarios [5]. By converting regression values to classification values, it can dramatically increase training speed and reduce the amount of memory used. Due to this, it is a much faster and lighter alternative to the XGBoost algorithm, which is why it is chosen for this study.

AdaBoost Ensemble Regressor (ABR). An AdaBoost regressor is a meta-estimator that begins by fitting another model on the original dataset and then fits additional copies of that model on the same dataset, but where the weights of instances are adjusted according to the error of the current prediction [1]. It creates more versions of the same model to tackle different sections of the training data, reducing error overall. Due to the large number of varying estimators that AdaBoost creates, the model is much less prone to overfitting than other models. The model we chose to train AdaBoost with is KNeighbors, which is described below.

KNeighbors Regressor (KNN). K-Nearest Neighbors (KNN) classifies a data point based on its nearest neighbors in the graph [14]. This algorithm is a non-parametric supervised learning method used for classification and regression.

In regression cases, the model takes the output value from a specific number of its nearest neighbors in the data, averages those values, and outputs that average. This algorithm does not make assumptions, so it handles outliers and minimizes error much better than decision trees and linear regression in many cases. This model was chosen for this study due to its novel approach to ensemble learning, high training speed, and high performance.

Elastic Net Regressor (ENR). Elastic Net is a regularized regression model that combines l1 and l2 penalties, i.e., lasso and ridge regression [34]. By combining both penalties, this model dramatically reduces overfitting. However, this model also performs feature selection, removing unnecessary features from the data. It was selected for this study because of its novel use of penalties and feature selection.

3 Datasets

3.1 Dataset of Spectroscopy Images

For this study, we used the non-invasive blood glucose monitor prototype ("Glucocheck") presented in our previous work [31]. Images were chosen instead of other forms of spectroscopy measurement, such as light intensity and PPG signals, because image capture is more replicable, accessible, and faster than other methods of spectroscopy data collection. Spectroscopy images were collected from the fingers of 43 participants between 18–65 years old. Two sets of 15 images were collected per participant. The first set was collected in a low-glucose fasting state, while the second set was collected one hour following a meal. Blood glucose was determined via finger prick using a commercial glucometer (FORA 6 Connect BG50) per manufacturer instructions. A set of 4 images is presented in Fig. 3. The images were taken after the finger prick at seconds 8 (top left), 16 (top right), 24 (bottom left), and 32 (bottom right). All images were collected from fingertips in the same format. A 640 × 480 resolution was chosen to preserve small details without sacrificing computing time and resources. The standard RGB color format was used. After removing any unclear images, the final dataset consisted of 1128 samples, each with two features, the image, and the corresponding blood glucose value.

3.2 Data Collection Ethics

The study was approved by the Institutional Review Board at Kennesaw State University (IRB-FY22-318). All participants provided written consent before participating.

3.3 Modified Datasets for CNN and Linear Models

Data transformation techniques were applied to the original data to generate three datasets, as described below.

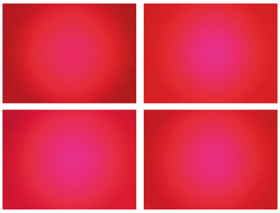


Fig. 3. Example of finger spectroscopy images collected from one individual.

Image Tensor Dataset. The "Tensor Dataset" was created in order to train the CNN models (VGG16 and MobileNetV2). Tensors are multi-dimensional matrices of numbers used in linear algebra; however, their application extends to images since images are multi-dimensional matrices of numbers as well. An image matrix consists of three dimensions: height, width, and color (red, green, and blue).

To convert an image into a tensor, a three-dimensional matrix (tensor) is created with the resolution of the image and the color format. Since images used in this study were 640 × 480 using the RGB color format, the image tensor was 640 pixels by 480 pixels by three colors. Then each color value for each pixel was entered into each value in the tensor, obtaining a tensor of 921,600 values. The resulting image tensor dataset was maintained at $160 \times 120 \times 3$ pixels to decrease computational time and necessary resources, when compared to a 640 × 480 dataset. The final dataset included the tensors with their corresponding blood glucose value. A visual demonstration of the image-tensor conversion can be seen in Fig. 4.

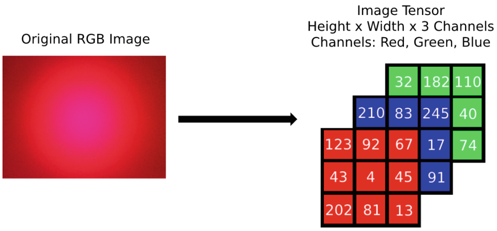


Fig. 4. Demonstration of Image Tensor Conversion (Color figure online)

CNN models are the only ones that can be trained with tensors because they use filtering techniques to analyze and process them. These filtering techniques are not available in other machine learning algorithms, which is why we used the two CNN models MobileNetV2 and VGG16.

Color Intensity Datasets. We have created four datasets based on extracting color intensity from the original images. For each possible value of red, green, and blue (0–255), the number of pixels with that same value in an image can be counted and recorded in a histogram [2]. Through this process, a histogram with RGB values on the x-axis (256 possible values for red, green, and blue) and the number of pixels on the y-axis can be created, as shown by Fig. 5. This process of counting pixel-intensity values for each color was used to create three datasets: “Red Intensity”, “Blue Intensity”, and “Green Intensity”. Each dataset consists of 257 features: 256 features for each possible value of that color and one feature for the blood glucose value of that image. Lastly, a final dataset, named “RGB-Intensity”, was created by combining the intensity values for all three-color channels. The RGB-Intensity dataset consisted of 769 features: 256 values of red, 256 values of green, 256 values of blue, and one value for blood glucose [2].

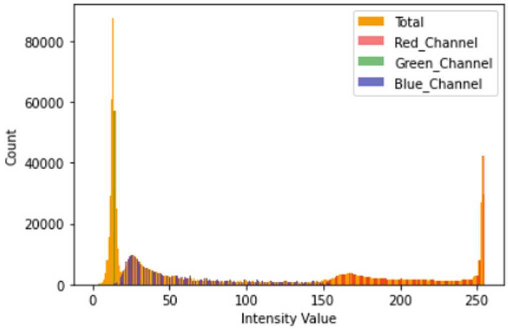


Fig. 5. Histogram of RGB Intensity Values in an Image. (Color figure online)

Image Measurement Datasets. The last five datasets were created by extracting measurement data from the images. To create the dataset, each image in the dataset was split into four channels: red, green, blue, and grayscale (the image with color removed). Then, for each color channel, the channel’s pixel center of mass, minimum, maximum, mean, median, standard deviation, and variance were calculated. To calculate these values: the images are first converted into numerical tensors, then their tensors (3-dimensional matrix) are converted into an array for each channel, and then each channel array (1-dimensional list) is used for calculations such as mean, median, minimum, maximum, etc. A demonstration of this process can be seen in Fig. 6.

Values for each channel were compiled into the same dataset with the correct blood glucose value and repeated for every image. The resulting “Measurement Dataset”, consisted of 29 features: seven measurements for each of the four channels and one feature for the blood glucose. After the creation of this dataset, four new datasets were created by merging the measurement features of each image with the intensity values of the same image created in the previously mentioned intensity datasets. This process resulted in four new datasets:

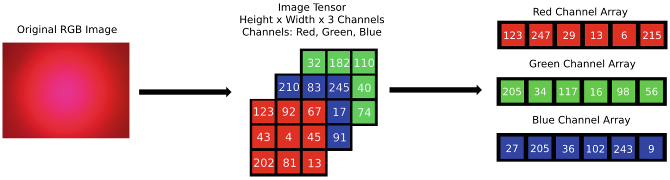


Fig. 6. Demonstration of Measurement Dataset Creation (Color figure online)

"Red-Measurement", "Green-Measurement", "Blue-Measurement", and "RGB-Measurement". The first three new datasets contained 285 features: 256 for the pixel intensities, 28 for the measurement features, and one for the blood glucose value. The last new dataset contained 797 features: 256 for each color channel, 28 measurement features, and one for the blood glucose value.

4 Experiment

After creating the datasets, each model was trained, tuned, and tested to each dataset to compare results, with only two exceptions. Since they can only be trained on tensor data, VGG16 and MobileNetV2 were only trained on the Tensor Dataset. Furthermore, the other linear models can only be trained on scalar data, so they were trained on every dataset except for the Tensor Dataset. The CNN models were trained using image data generators, which come with the TensorFlow library for Python that was used for training models. Moreover, before the training process, the image data generators were used to scale down the pixel values from 0–255 to 0–1 to reduce error and GPU usage. Besides these changes during training, the testing of CNN models was the same as the other models. On another note, since the AdaBoost Ensemble Learning algorithm uses another algorithm as a base estimator, for each dataset, the AdaBoost model was trained with the model that had the highest accuracy for that dataset. A summary of the models trained with each dataset can be seen in Table 1.

Table 1. Models Trained on Each Dataset

Tensor Dataset	Intensity Dataset	Measurement Dataset	Intensity-Measurement Dataset
VGG16	Random Forest	Random Forest	Random Forest
MobileNetV2	Support Vector	Support Vector	Support Vector
	Bayesian Ridge	Bayesian Ridge	Bayesian Ridge
	XGBoost	XGBoost	XGBoost
	HGB	HGB	HGB
	AdaBoost	AdaBoost	AdaBoost
	KNeighbors	KNeighbors	KNeighbors
	Elastic Net	Elastic Net	Elastic Net

4.1 Training and Hyperparameter Tuning of Models

To train the models, all of the datasets were split into training/testing splits where the training data was used to fit the model, and the testing data was used to measure the model's performance. The training/testing split ratio was 75:25 to ensure sufficient data to train the models and ensure that they would not overfit. After creating training/testing splits, each model in the set was fitted to training data and then tested. However, to ensure that the models were compared effectively, each model's hyperparameters were tuned to each specific dataset to minimize error and overfitting. After the models were finished with training and tuning, they were tested, and the results were recorded in a table.

4.2 Testing Models

Three distinct metrics were considered for testing/tuning the models: MAE, RMSE, and the Clarke Error Grid. The MAE is the mean of all errors between the blood glucose values that a model predicts and the actual blood glucose value tied to an image. The RMSE is the root of the mean of each error squared. MAE is a more direct metric for calculating error as it is unbiased towards all errors and treated as an average. However, because it squares the errors, RMSE is biased against large prediction errors, making it weighted against outliers. RMSE is usually used in scenarios when an increase in error is disproportionate to the effect, for example, if the error increases from 5 to 10 and the effect is four times as bad. Since RMSE is always higher or equal to MAE, the difference between the two values is critical for evaluating outliers. If RMSE is significantly higher than MAE, then there are outliers in the predictions. For this reason, RMSE was used to tune the models to reduce overfitting but not recorded in the results or evaluation. Lastly, Clarke Error Grids were used to evaluate models since they have been widely used for several decades to evaluate the performance of blood glucose meters. Clarke Error Grids are scatterplots with predicted blood glucose values on the y-axis and actual blood glucose values on the x-axis. The

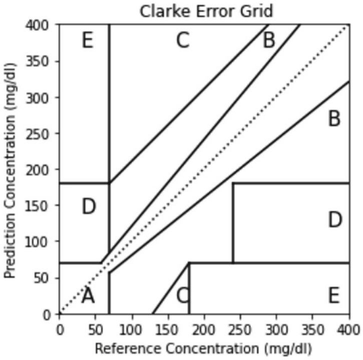


Fig. 7. Clarke Error Grid

grid is split into several zones, and each zone signifies a level of risk of a negative outcome due to the measurement error in blood glucose values which can be seen in Fig. 7.

There are 5 zones: A - Accurate, B - Clinically Acceptable, C - Overcorrection, D - Failure to Detect/Treat, and E - Erroneous Treatment [8]. These three metrics were all used when measuring the performance of the models during training and testing.

5 Evaluation

For comparing the performance of the models we used MAE and Clarke Error Grid (Zone A Percentage) metrics. The percentage of data points that fall into each zone of the clinical outcome can be determined by analyzing the grid. To get Zone A Percentage, the number of predictions in Zone A (Clinically Accurate) is recorded as a percentage of the total number of predictions made. After the models were trained and tuned, they were tested with the testing data, and the results were recorded in Table 2 and Table 3 respectively.

Table 2. Model Testing Results from Tensor and Intensity Datasets - MAE and Zone A percentages from clarke error grid analysis

	Image Tensor (IT)	Red-Intensity (RI)	Green-Intensity (GI)	Blue-Intensity (BI)	RGB-Intensity (RGBI)
VGG16	16.58–87.59%	–	–	–	–
MobileNetV2	15.68–87.23%	–	–	–	–
Random Forest	–	13.17–86.17%	13.31–85.11%	14.04–86.17%	12.46–88.65%
Elastic Net	–	15.59–85.46%	16.23–82.27%	15.53–84.4%	14.42–84.4%
KNeighbors	–	9.88–90.78%	14.06–88.3%	14.35–85.46%	10.84–88.65%
Support Vector	–	14.43–89.36%	15.71–89.36%	14.3–89.36%	13.14–88.65%
Bayesian Ridge	–	15.43–85.11%	16.01–83.33%	15.34–83.33%	14.28–84.4%
XGBoost	–	12.93–87.94%	14.1–84.75%	13.97–84.75%	12.26–89.72%
HGB	–	13.12–86.88%	14.99–84.04%	14.37–83.69%	12.53–87.59%
AdaBoost	–	9.66–90.78%	13.31–87.94%	14.08–85.46%	10.95–88.65%

Table 3. Model Testing Results from Measurement Datasets - MAE and Zone A percentages from clarke error grid analysis

	Measurement (ME)	Red-Measurement (RM)	Green-Measurement (GM)	Blue-Measurement (BM)	RGB-Measurement (RGBM)
Random Forest	14.27–83.33%	12.85–87.23%	12.63–86.52%	13.91–85.82%	12.74–88.65%
Elastic Net	16.38–81.56%	15.68–84.04%	16.89–85.11%	15.55–81.56%	14.41–83.69%
KNeighbors	15.02–81.91%	9.55–90.78%	14.3–86.17%	15.81–84.4%	12.43–87.59%
Support Vector	16.13–89.72%	14.3–87.94%	15.02–89.01%	14.58–87.23%	13.28–87.94%
Bayesian Ridge	16.37–81.56%	15.52–84.04%	17.43–85.46%	15.52–82.62%	14.3–83.33%
XGBoost	14.51–83.69%	13.03–86.88%	12.86–88.3%	13.6–86.52%	12.89–87.59%
HGB	14.78–81.21%	13.71–85.11%	13.17–86.17%	13.7–85.11%	12.58–87.94%
AdaBoost	15.13–80.5%	9.4–90.78%	12.74–86.88%	13.41–87.59%	13.18–86.52%

6 Discussion

AdaBoost with KNeighbors trained on the Red-Measurement dataset provided the most accurate estimates of blood glucose among all of the dataset-models tested. This dataset-model combination had an MAE of 9.4 mg/dl, an RMSE of 16.72 mg/dl, and a Clarke Error Grid Zone A Percentage of 90.78% illustrated in Fig. 8.

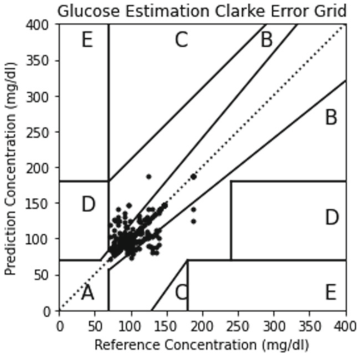


Fig. 8. Clarke Error Grid of AdaBoost model with KNeighbors Trained on Red-Intensity Dataset

From best to worst, the models ranked AdaBoost, KNeighbors, Random Forest, XGBoost, HGB, Support Vector, Bayesian Ridge, Elastic Net, MobileNetV2, and VGG16 as displayed in Fig. 9. From best to worst, the datasets ranked RGB Intensity, Red Measurement, Red Intensity, RGB Measurement, Green Measurement, Blue Intensity, Blue Measurement, Green Intensity, Measurement, and Image Tensor as shown in Fig. 10. The datasets containing Red and RGB data outperformed the other datasets by a large margin. However, combining measurement and intensity values did not seem to improve performance for the red dataset, but instead hindered it. Datasets with Blue and Green data appeared to perform equally, but their performance was inferior to the Red and the combined RGB overall. Furthermore, the Green data, but not the blue, seemed to perform better after combining intensity and measurement data. The intensity datasets performed better than the measurement datasets, and the dataset with only measurement values performed significantly worse. The image tensor dataset performed the worst of all datasets, while the CNN models performed the worst among the group of models. AdaBoost and KNeighbors performed the best with every dataset they were trained on, while XGBoost, Random Forest, and HGB generally outperformed the other models. These results suggest that the best data for blood glucose estimation by spectroscopy is color intensity data focused on either the red channel or all three channels. The results further suggest that the KNeighbors algorithm is well-suited for blood glucose estimation with scalar data, and using AdaBoost as an ensemble learner can boost performance. Models that use boosting and bagging (XGBoost, AdaBoost, HGB, etc.) outperformed models that do not (Elastic Net, Bayesian Ridge, Support Vector). Furthermore, the penalties and feature selection in Elastic Net and the binning in Histogram-Based Gradient Boosting did not seem to increase performance compared to bagging and boosting. Finally, both the dataset and model results suggest that Convolutional Neural Networks and Tensor datasets perform worse than Linear Models, Ensemble Learners, and Scalar Data.

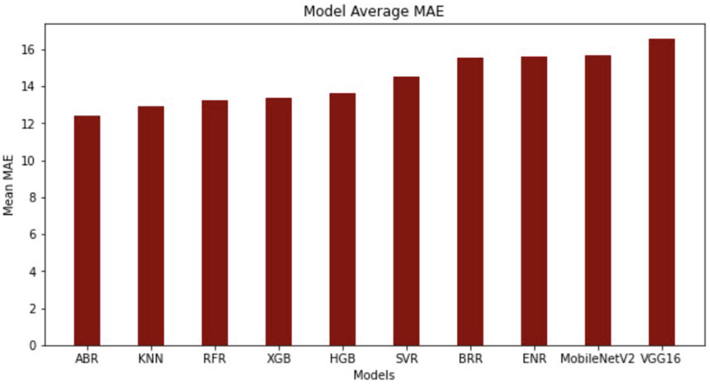


Fig. 9. Model Average MAE

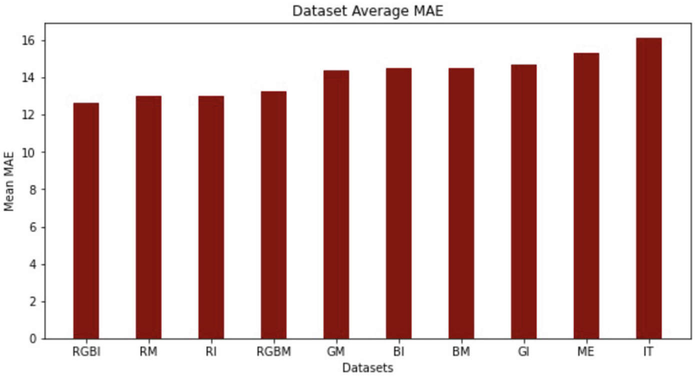


Fig. 10. Dataset Average MAE

7 Conclusion

From training, tuning, and testing ten machine learning models on ten different datasets, we have determined that the best model for estimating blood glucose through spectroscopy images is AdaBoost trained with KNeighbors. Furthermore, the best image data to train the model is color intensity data collected from the red channel. Our highest performing dataset and model recorded a final Mean Absolute Error of 9.4, a Root Mean Squared Error of 16.72, and a Clark Error Grid Zone A Percentage of 90.78%. We also showed that intensity data outperformed measurement and tensor data, while the red and RGB channels outperformed all other color channels. Furthermore, models that utilize bagging and boosting outperformed those which did not, while linear models outperformed CNN models, regardless of their support for bagging or boosting.

References

1. Sklearn.ensemble.adaboostregressor. https://scikit-learn.org
2. Alarcón-Paredes, A., Francisco-García, V., Guzmán-Guzmán, I.P., Cantillo-Negrete, J., Cuevas-Valencia, R.E., Alonso-Silverio, G.A.: An IoT-based non-invasive glucose level monitoring system using Raspberry Pi. Appl. Sci. **9**(15), 3046 (2019). https://www.mdpi.com/2076-3417/9/15/3046/htm
3. Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET), pp. 1–6. IEEE (2017)
4. Amir, O., et al.: Continuous noninvasive glucose monitoring technology based on "occlusion spectroscopy" (2007)
5. Brownlee, J.: Histogram-based gradient boosting ensembles in Python (2021). https://machinelearningmastery.com/
6. Brownlee, J.: XGBoost for regression (2021). https://machinelearningmastery.com/xgboost-for-regression/
7. Centers for Disease Control and Prevention (CDC): National Diabetes Statistics Report website (2018). https://www.cdc.gov/diabetes/data/statistics-report/index.html. Accessed 2022
8. Clarke, W.L., Cox, D., Gonder-Frederick, L.A., Carter, W., Pohl, S.L.: Evaluating clinical accuracy of systems for self-monitoring of blood glucose (1987). https://doi.org/10.2337/diacare.10.5.622
9. Donges, N.: Random forest algorithm: a complete guide. https://builtin.com/data-science/random-forest-algorithm
10. Enejder, A.M., et al.: Raman spectroscopy for noninvasive glucose measurements. J. Biomed. Opt. **10**(3), 031114 (2005)
11. Haxha, S., Jhoja, J.: Optical based noninvasive glucose monitoring sensor prototype. IEEE Photonics J. **8**(6), 1–11 (2016)
12. Hull, E.L., et al.: Noninvasive skin fluorescence spectroscopy for detection of abnormal glucose tolerance. J. Clin. Transl. Endocrinol. **1**(3), 92–99 (2014)
13. Kasahara, R., Kino, S., Soyama, S., Matsuura, Y.: Noninvasive glucose monitoring using mid-infrared absorption spectroscopy based on a few wavenumbers. Biomed. Opt. Express **9**(1), 289–302 (2018)
14. Kramer, O.: K-nearest neighbors. In: Kramer, O. (ed.) Dimensionality Reduction with Unsupervised Nearest Neighbors, pp. 13–23. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38652-7_2
15. Maruo, K., et al.: Noninvasive blood glucose assay using a newly developed near-infrared system. IEEE J. Sel. Top. Quantum Electron. **9**(2), 322–330 (2003)

16. Moore, J.X., Chaudhary, N., Akinyemiju, T.: Peer reviewed: metabolic syndrome prevalence by race/ethnicity and sex in the United States, National Health and Nutrition Examination Survey, 1988–2012. Preventing Chronic Dis. **14** (2017)
17. Noble, W.S.: What is a support vector machine? Nat. Biotechnol. **24**(12), 1565–1567 (2006)
18. Pai, P.P., Sanki, P.K., Sahoo, S.K., De, A., Bhattacharya, S., Banerjee, S.: Cloud computing-based non-invasive glucose monitoring for diabetic care. IEEE Trans. Circuits Syst. I Regul. Pap. **65**(2), 663–676 (2017)
19. Pickup, J.C., Khan, F., Zhi, Z.L., Coulter, J., Birch, D.J.: Fluorescence intensity- and lifetime-based glucose sensing using glucose/galactose-binding protein. J. Diab. Sci. Technol. **7**(1), 62–71 (2013)
20. Pitzer, K.R., et al.: Detection of hypoglycemia with the GlucoWatch biographer. Clin. Diabetol. **2**(4), 307–314 (2001)
21. Rachim, V.P., Chung, W.Y.: Wearable-band type visible-near infrared optical biosensor for non-invasive blood glucose monitoring. Sens. Actuators B Chem. **286**, 173–180 (2019)
22. Raj, A.: Unlocking the true power of support vector regression (2020)
23. Robinson, M.R., et al.: Noninvasive glucose monitoring in diabetic patients: a preliminary evaluation. Clin. Chem. **38**(9), 1618–1622 (1992)
24. Rothman, A.: The Bayesian paradigm & ridge regression (2020). https://towardsdatascience.com
25. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
26. Saklayen, M.G.: The global epidemic of the metabolic syndrome. Curr. Hypertens. Rep. **20**(2), 1–8 (2018)
27. Sakr, M.A., Serry, M.: Non-enzymatic graphene-based biosensors for continous glucose monitoring. In: 2015 IEEE SENSORS, pp. 1–4. IEEE (2015)
28. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
29. Shi, B., et al.: Learning better deep features for the prediction of occult invasive disease in ductal carcinoma in situ through transfer learning, p. 98 (2018). https://doi.org/10.1117/12.2293594
30. Tammina, S.: Transfer learning using VGG-16 with deep convolutional neural network for classifying images. Int. J. Sci. Res. Publ. (IJSRP) **9**(10), 143–150 (2019)
31. Valero, M., et al.: Development of a non-invasive blood glucose monitoring system prototype: pilot study. J. Med. Internet Res. JMIR Formative Res. (forthcoming/in press)
32. Vashist, S.K.: Non-invasive glucose monitoring technology in diabetes management: a review. Anal. Chim. Acta **750**, 16–27 (2012)
33. Vegesna, A., Tran, M., Angelaccio, M., Arcona, S.: Remote patient monitoring via non-invasive digital technologies: a systematic review. Telemed. e-Health **23**(1), 3–17 (2017)
34. Verma, Y.: Hands-on tutorial on elasticnet regression (2021). https://analyticsindiamag.com/hands-on-tutorial-on-elasticnet-regression/
35. Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. Fron. Comput. Sci. **14**, 241–258 (2020). https://doi.org/10.1007/s11704-019-8208-z

Classification of Kidney Tumor Grading on Preoperative Computed Tomography Scans

Maryamalsadat Mahootiha[1,2](✉), Hemin Ali Qadir[1], Jacob Bergsland[1], and Ilangko Balasingham[1,3]

[1] The Intervention Centre, Oslo University Hospital, 0372 Oslo, Norway
marymaho@uio.no
[2] Faculty of Medicine, University of Oslo, 0372 Oslo, Norway
[3] Department of Electronic Systems, Norwegian University of Science and Technology, 7491 Trondheim, Norway
https://www.ous-research.no/interventionalcentre/

Abstract. Deep learning (DL) has proven itself as a powerful tool to capture patterns that human eyes may not be able to perceive when looking at high-dimensional data such as radiological data (volumetric data). For example, the classification or grading of kidney tumors in computed tomography (CT) volumes based on distinguishable patterns is a challenging task. Kidney tumor classification or grading is clinically useful information for patient management and better informing treatment decisions. In this paper, we propose a novel DL-based framework to automate the classification of kidney tumors based on the International Society of Urological Pathology (ISUP) renal tumor grading system in CT volumes. The framework comprises several pre-processing techniques and a three-dimensional (3D) DL-based classifier model. The classifier model is forced to pay particular attention to the tumor regions in the CT volumes so that it can better interpret the surface patterns of the tumor regions to attain performance improvement. The proposed framework achieves the following results on a public dataset of CT volumes of kidney cancer: sensitivity 85%, precision 84%. Code used in this publication is freely available at: https://github.com/Balasingham-AI-Group/Classification-Kidney-Tumor-ISUP-Grade.

Keywords: Kidney cancer · Renal cancer · Deep neural networks · Tumor grading · Classification · CT scan

1 Introduction

Kidney cancer (or renal cancer) is among the most commonly diagnosed visceral malignancies, with a significant annual increase in the incidence- and mortality-rate accounting for 431,288 new cases and 179,368 new deaths in both genders

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 75–89, 2023.
https://doi.org/10.1007/978-3-031-34586-9_6

in 2020 [1]. Surgical removal is still the most common treatment option for localized kidney tumors. Recently, several other targeted therapies for the treatment of kidney cancer have been introduced to improve patient outcomes and avoid surgical intervention [2,3]. Accurate grading and classification of renal cell neoplasia are essential to provide the optimal treatment option and play a major role in the estimation of patient prognosis. There are several grading systems, with Fuhrman grading being the most widely used one. Recently, there have been doubts about the applicability and prognostic value of Fuhrman grading [4]. In 2012, the International Society of Urological Pathology (ISUP) held a conference to address these issues and proposed a novel grading system known as ISUP grading classification, categorizing renal cell carcinoma (RCC) into four grades namely grades 1, 2, 3, and 4 [5]. It has been shown that a tumor's specific information can be observed pre-operatively from the tumor's appearance on cross-sectional imaging such as computed tomography (CT) scan or magnetic resonance imaging (MRI) [6]. Manual interpretation and quantitative evaluation of radiological data is a laborious and noisy process. In addition, there can be hidden information that the human brain can not perceive from this type of data. For example, microscopic morphological changes associated with histological patterns are crucial in establishing the ISUP grading system.

Over the last decade, several computational methods have been proposed to automate renal cancer classification and staging [7–10]. Deep learning (DL) has been the dominant method because of its advances in finding complex hidden patterns from training data and transforming the input images into abstract features. In most of the studies [7–10], renal whole-slide histology images have been the major source of information about microscopic morphological patterns which are associated to different RCC subtypes such as clear cell RCC, papillary RCC, chromophobe RCC, renal oncocytoma, etc. In contrast, there are several attempts to utilize radiological data for the development of automatic kidney cancer classification [11–15] and staging [16,17]. Many DL-based models were proposed for binary classification differentiating benign and malignant renal tumors from either CT scans [13,15] or MRI [12,14]. S. Han [11] modified GoogleNet [16] for discriminating three major subtypes of RCCs using CT image analysis. N. Hadjiyski et al. [16] adapted the 3D variant of the inception model to predict cancer staging, while M. A. Hussian et al. [17] proposed an automatic low stage (I-II) and high stage (III-IV) RCC classification both from CT scans.

In this paper, we propose a novel DL-based framework to computationally classify kidney tumors into ISUP grades from pre-operative CT scans available for each patient. To the best of our knowledge, our work is the first study to investigate DL in 3D images for renal tumor ISUP grading indicating the histopathological patterns and associated with risk score [5,18], which is the basis of the survival analysis. We do not intend to detect and segment the kidneys and the tumors in CT volumes; instead, we assume that the kidneys and tumors are already localized and segmented. We first extract the kidneys from a 3D CT volume using the provided manual ground truth of kidneys. We concatenate the extracted kidneys and the corresponding ground truth of the tumors into a single tensor on the channel dimension. This concatenation step aims to force the DL-based classifier model to pay particular attention to the surface patterns of the tumor regions. The concatenated tensor is then fed into a three-dimensional (3D) convolutional neural network (CNN) to classify the kidney image(s) into 4 ISUP grades in every CT volume. We adapt EfficientNet [19] as our classifier model. More specifically, we transform the 2D EfficientNet-B7 to its 3D variant so that it can handle 3D volumetric data. We apply various data augmentations to overcome the class imbalance issue and feed the training model with more samples and several preprocessing methods to standardize the inputs and improve their quality. We show that our proposed framework can provide promising results on unseen CT volumes. This initial result can further be developed into a prognosis model, survival analysis, and treatment management plan.

2 Related Work

Recent studies have focused on the automated grading of clear cell renal cancer carcinoma (ccRCC). Some of them employed histopathology images, while others used CT or MRI for this prediction.

For histopathology images, Tian et al. [20], and Yeh et al. [21] employed Fuhrman's grading system to identify ccRCC as low or high grade. In both studies, the authors, in collaboration with pathologists, examined whole-slide images, identified regions of interest (ROIs), and assigned a grade to each ROI. Then, features of histopathology images were retrieved from ROIs for the model training. Tian et al. tried to find an optimal way between a neural net, random forest, and support vector machine for the model training, while Yeh et al. tried to use a support vector machine to train the classifier model. Both Tian et al. and Yeh et al. could get high sensitivity, specificity, and AUC for their models, and they recommended predicting ISUP grades in future work.

For radiological imaging, Sun et al. [22] developed the support vector machine-based method to determine the ISUP grade of kidney cancer with clear cells. In this research, CT images were divided into two categories: low and high grade. Resampling and a Gaussian filter were utilized for denoising at the preprocessing level. Then, the greatest cross-section picked by radiologists was utilized as the ROI. Sun et al. used a feature extraction mechanism to generate three distinct predictive models, each of which was based on a distinct selection of features. The AUC for the third model was the highest at 0.91. Another research study led by Cui [23] graded ccRCC using CT and MR based on the decision tree. Normalization and pixel resampling were utilized for preprocessing level in this research. The classification was determined by low and high ISUP grades. The RoI was determined based on the tumor-containing slices. Cui et al. next attempted to extract the texture of the slices and create features from them; they employed a decision tree to predict a low or high ISUP grade and attempted to test the model using ACC. For Cui et al. model, an ACC greater than 0.70 was achievable. In a different study, Zhao et al. [24] classified MRI images based on ISUP and Fuhrman grading as low or high grade using CNN. This study was a binary classification: low and high grade. Data augmentation was utilized prior to the model training. The model with the highest AUC was selected as the ultimate model. The model was developed using ResNet50 and 2D CNN. Zhao et al. combined t1 and t2 sequences for the model and included clinical variables such as age, gender, and tumor size in the network design. For low and high ISUP classification, they could gain 0.92 in sensitivity and 0.78 in specificity. These studies [22–24] agree that CT texture analysis can predict ccRCC pathologic nuclear grade noninvasively.

Multiple factors make our methodology superior to that of previous research studies. First, it is based on CT images, which is a non-invasive method; second, it uses deep learning and does not require feature extraction; third, it uses 3D images and 3D models for predicting, so we do not lose any information by changing it to a 2D based model; fourth, we attempted to have four output classifications rather than a binary classification; and finally, we do not employ clinical data in addition to the CT images, therefore our prediction is solely based on the CT scans and does not require any further information.

3 Methods and Materials

Figure 1 illustrates our proposed DL-based framework developed for kidney tumor grading classification based on the ISUP grading system. Every step will be explained in detail in the following sections.

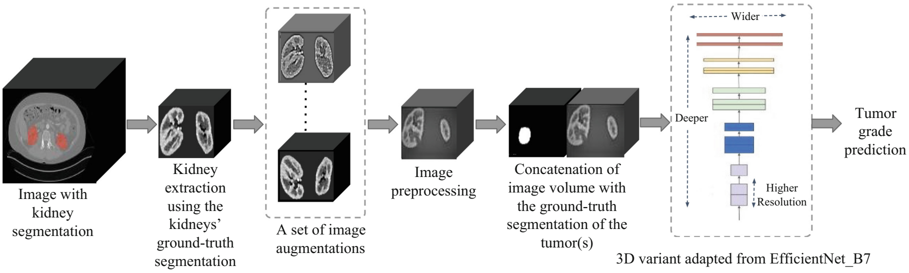


Image with kidney segmentation
Kidney extraction using the kidneys' ground-truth segmentation
A set of image augmentations
Image preprocessing
Concatenation of image volume with the ground-truth segmentation of the tumor(s)
Wider
Deeper
Higher Resolution
3D variant adapted from EfficientNet_B7
Tumor grade prediction

Fig. 1. Overview of the framework proposed for kidney image classification based on the ISUP grading system. We separate the left and right kidneys from the whole image slices during image preparation. We enlarge the number of samples in the training dataset by applying various forms of data augmentation strategies. We enhance the data quality by improving image quality, resizing, and re-orienting the volumes in the image pre-processing phase. To force the model to focus on the tumor surface patterns, we concatenate the image and manual segmentation of the tumors, and finally, we train the classification model with concatenated volumes. The model produces probability values for the four different ISUP grades as the output decision.

3.1 Classifier Architecture

The state-of-the-art convolutional neural network (CNN)architecture for image classification is called Efficient-Net [19]. In a quick but efficient way, Efficient-Net scales up models using the compound coefficient method. The authors of

EfficientNet proposed seven models of various dimensions, which exceeded the state-of-the-art accuracy of most CNNs and had a far higher degree of efficiency. The largest Efficient-Net model, Efficient-Net B7, obtained the best performance on the ImageNet and the CIFAR-100 datasets. The number of parameters in Efficient-Net B7 is higher than the other variants (e.g., B0, B1, B2, B3, B4, B5, and B6). In this study, we adapt the exact structure of Efficient-Net B7 and transform it to a three dimensional (3D) CNN model so that it can handle 3D image data such as CT volumes.

3.2 Dataset

In this paper, we use KiTS21 dataset [25] for training and testing our proposed method. This dataset consists of 300 different patients, each with clinical data and a CT scan with manually annotated kidneys and tumors (ground-truth labels). Patients receiving a partial or complete nephrectomy for suspected kidney cancer between 2010 and 2020 at either the M Health Fairview or Cleveland Clinic medical facility have been included in this dataset. Before surgery, all patients underwent a contrast-enhanced CT scan showing both kidneys. The primary purpose of gathering this dataset was to apply segmentation algorithms.

We attempt to use this dataset since it has a detailed clinical dataset, precise annotation, and adequate subjects. The dataset contains three files as follows: CT scan volumes, annotation volumes, and clinical data. All of the images are in NIFTI format. Each annotation volume contains manual segmentation of the kidneys, tumor(s), and cyst(s). Clinical data is in a JSON file format with 63 fields of clinical parameters for every patient. All essential clinical information, like pathology results, is stored in this file [26]. Originally this data came from the Cancer Imaging Archive, where the imaging and segmentation were stored in DICOM format, and the clinical data was a single CSV file[1].

3.3 Data Preparation

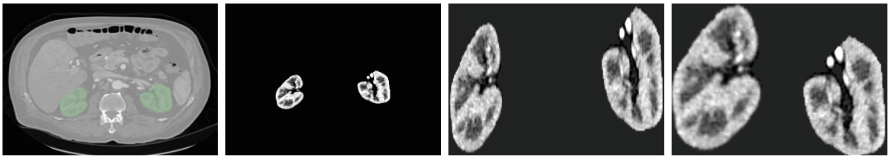
Data preparation is a pre-processing mechanism to structure, manipulate, and organize raw data to the data format that the training model can analyze more efficiently. In this study, we apply data preparation on the CT scan volumes and their corresponding annotation volumes.

Image Preparation. The 3D image data from the KiTS21 dataset depict the whole abdomen. The kidneys with tumors cover only a small percentage of the entire image slices. In this study, we aim to train our proposed framework to view only the imaging information related to the kidneys and the tumors. Therefore, we extract the left and right kidneys from the image volumes using the provided ground-truth annotation. Figure 2 shows the steps used to prepare the training samples by removing other organs and extracting the two kidneys from whole image volumes.

---

[1] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=61081171.

The image and segmentation volumes are stored as 3D arrays. The numbers in the image arrays are between 0 and 255, where 0 is the darkest part and 255 is the brightest part. The numbers in the segmentation arrays are 0, 1, 2 and 3 (1 = kidney, 2 = tumor, and 3 = cyst). For image preparation, we have 3 phases. In phase 1, we keep the numbers in image arrays that the corresponding segmentation is one and change the other numbers in image arrays to zero. This step will change all image space except two kidneys to black. Then in phase 2, we keep two kidneys and delete the black background. In phase 3, the black space between two kidneys should also be eliminated, as our first purpose was not to enter the black space into the training model. So we extract the left and right kidneys and merge them again in the width dimension in phase 3.



(a) KiTS21 Image (b) output of phase 1 (c) output of phase 2 (d) output of phase 3

Fig. 2. Image preparation process

Label Preparation. From the clinical dataset file that comes with the KiT21 dataset, we use the tumor_isup_grade field as the label for image classification. This clinical parameter has four values: 1, 2, 3, and 4. ISUP grade of the tumors was indicated in the post-operative surgical pathology report. We notice that in 56 cases, the value of Null is used where ISUP grade does not apply, such as benign tumors or Chromophobes. We remove those 56 patients from our training and testing dataset, leaving us with 244 samples in our final dataset.

3.4 Data Augmentation

The deep learning models frequently need a large amount of training data, which is not always available, in order to make accurate predictions. We apply data augmentation to increase the number of samples in the training dataset. After eliminating those patients without ISPU values, we are left with 244 samples. Patients with the ISUP1 class make 13% of the total, the ISUP2 class 48%, the ISUP3 class 27%, and those with the ISUP4 class 12%. This class imbalance leads to biasing impact on the model training and the final results—the trained model will be more biased toward the dominant class in the training dataset and show poor performance on the minor class. Another challenge that we encounter is that we have to train our classifier model from scratch as we are unable to apply transfer learning or fine-tune a pre-trained 3D EfficientNet-B7 transformed from the original 2D EfficientNet-B7 in this study. Two hundred forty-four samples might not be enough for training a deep neural network for an image classification model from scratch. A huge quantity of labeled training images is needed for deep learning models to be trained from scratch. We try to partially overcome these two problems with data augmentation.

When strategies like undersampling, oversampling, and data augmentation are used to fix the class balance issue, the model's efficacy increases [27,28]. We don't use oversampling as this method can lead to the model being overfitted to the minority class [28]. Additionally, we avoid using undersampling since we lack sufficient samples in the dataset and don't want to lose any data. As we can see in the literature, performance progress slowed down after 150 images in each class, and after 500 images in each class, there was no noticeable improvement [29]. We found that 500 images per class are enough to attain a reasonable classification accuracy. We increase the number of samples to 2000, 500 in each class. We calculate the number of subjects in each class and realize that class 1 would need to be augmented eleven times, class 2 twice, class 3 five times, and class 4 fourteen times.

For data augmentation, we do not employ generative adversarial networks and would rather use traditional approaches. The critical point is that if we want to do the augmentation for class4 fourteen times, we must make fourteen different augmented versions of the original data. We use MONAI transformers for data augmentation because the MONAI module is a comprehensive python library for manipulating 3D data such as volumetric images. MONAI library contains all recommended image augmentation techniques to enlarge the number of training samples. Table 1 shows the various transformations we use for data augmentation. We utilize the various combinations of transformers from Table 1 for data augmentation. Figure 3 displays one slice of the original patient's data along with three augmented versions of that slice.

Table 1. MONAI transformers used for data augmentation

Position Augmentation	Noise Augmentation
Affine	GaussianNoise
Rotate90	GaussianSmooth
Flip	GaussianSharpen
	GibbsNoise
	SpaceSpikeNoise

3.5 Data Splitting

Our augmented dataset consists of various copies of the original samples. To prevent unfair performance evaluation of our proposed framework, we split the dataset based on the patient ID into training and testing subsets. In this way, we avoid having the same patient with all its augmented versions in both the training and testing subsets. We use the K-fold cross-validation technique to split out the dataset. We use 3-fold cross-validation. We split our dataset randomly into three

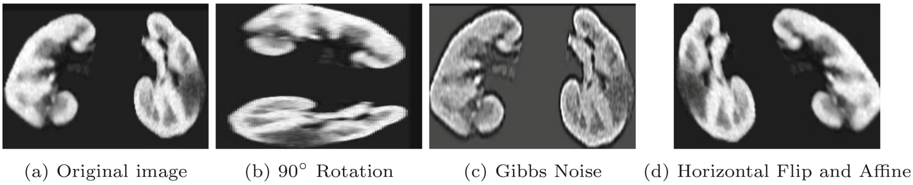

(a) Original image (b) 90° Rotation (c) Gibbs Noise (d) Horizontal Flip and Affine

Fig. 3. Comparison between one axial slice of original image with 3 different augmented versions

different subsets: 162 samples in the training subset; 82 samples in the testing subset. We choose 10% (16 samples) of the training subset as our validation subset in every fold. As we intend to have the same ISUP class distribution in the validation subset, we select four samples from each ISUP grade class.

3.6 Image Pre-processing

Image pre-processing is an essential step before image classification. The purpose of pre-processing is to enhance the image's quality and modify a few of its features so that the training model can better interpret the input [30,31]. We resize all the volumes to 128 × 128 × 128 to have the same size volumes for training the model. We follow the recommended size by the MIT challenge[2] to make the data more manageable. We do not select a bigger size like 256 because a larger image resolution is expensive both in terms of computational power and memory [30]. One millimeter isotropic voxel size is chosen for every volume. This is the standard voxel size recommended by previous studies [32,33]. We re-orient all volumes to the RAS (Right, Anterior and Superior). This is the most common orientation used in medical images [32–34]. We use intensity normalization based on the Z-score in medical imaging [30,35]. We use the image contrast part in ITK snap software[3] for this normalization. We showed the images to the clinicians to identify which contrast range between the kidney and the tumor was more noticeable. So we can figure out the minimum and the maximum contrast number in which the tumor is more distinctive from the kidney. We change intensity values in the image arrays based on this image contrast range.

For kidney image and tumor segmentation, we utilize identical image pre-processing transformers; however, we do not apply intensity normalization for tumor segmentation because the contrast of the segmentation image is not important for training the model.

3.7 Kidney and Tumor Concatenation

In this study, our goal is to classify kidney tumors based on distinguishable surface patterns. To force our 3D EfficientNet-B7 to pay particular attention to

---

[2] http://6.869.csail.mit.edu/fa17/miniplaces.html.

[3] http://www.itksnap.org/pmwiki/pmwiki.php.

the surface patterns on the tumors, we concatenate the extracted kidneys with their corresponding provided manual segmentation of the tumors. In addition, this image concatenation enriches the input volume with the location and size of the tumors. If we train our 3D EfficientNet-B7 on the kidneys only without providing the location of the tumors, the model may look at other parts of the input volumes and find other patterns and associate them to the classes. This leads to poor performance on unseen data.

3.8 Training Details

We use the Pytorch library for training our model. The experiments are executed in the Linux Ubuntu Operating system on a machine with AMD Ryzen 7 5800X 8-Core Processor, NVIDIA GeForce RTX 3090 GPU and 32 GB RAM. Based on the three folds we previously acquired, we train our model three times but with the same hyperparameters. Each time validation set contains around 10% of the training set. During training, none of the samples from the validation sets are utilized to determine the loss function or back-propagate gradients across the network.

After every training epoch, the model is evaluated on the complete validation set, and the mean AUC[4] is calculated. Model parameters are stored, overwriting the previous model, each time a new best mean validation AUC is obtained. In this regard, compared to all training epochs, the final model that is created during training has the greatest mean validation AUC. We decide on 50 as the number of epochs since we see that the training losses stop decreasing after about 50 epochs. Each model is trained with the help of the Cross-Entropy loss, which is given by:

$$L = -\sum_{i=1}^{n} t_i \times log(p_i), \tag{1}$$

where t_i is the true label and p_i is the softmax probability for the ith class and n is the number of classes.

The ADAM optimizer [36] is used to train the models, and a learning rate of 1×10^{-4} is used since it is empirically proven to produce the best results on clean data [37]. Ten batches are selected to train the model based on the image sizes and computing memory.

4 Results and Discussion

After training the model on the three folds, we evaluated the model's performance. Precision, Recall, and F-score metrics were used to quantitatively evaluate the performance of the proposed framework. The performance metrics are computed from the following formulas:

---

[4] Area under the curve is a performance measurement for the classification problems. It tells how much the model is capable of distinguishing between classes.

$$\textbf{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{2}$$

$$\textbf{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{3}$$

$$\textbf{F-score} = 2\times\frac{\text{Precision } \times \text{ Recall}}{\text{Precision} + \text{Recall}}, \tag{4}$$

TP is the number of samples that are truly classified, FP is the number of samples that should be in an ISUP class except for the ISUP class-x, but they belong to ISUP class-x; and FN is the number of samples that should be in ISUP class-x, but they are in the other ISUP classes.

Precision, Recall, and F-score was computed for each ISUP class. We calculated the average of four Precision, Recall, and F-scores we gained for each ISUP class. We repeated this process three times for each of the three folds we had, giving us three average Precision, Recall, and F-scores. For our model, we obtained a total Precision of 0.74, Recall of 0.71, and F-score of 0.72 by calculating the mean three average Precision, Recall, and F-scores. Table 2 shows the performance metrics in fold two in which the best performance was obtained.

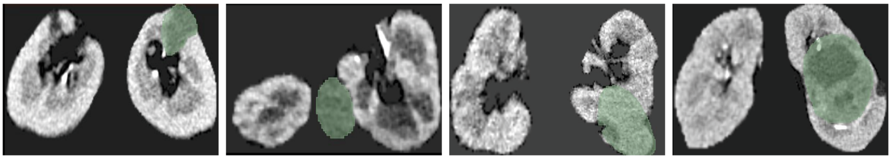
Table 2. Fold 2 performance evaluation of the proposed framework

	Precision	Recall	F-score
ISUP1 class	0.86	0.91	0.88
ISUP2 class	0.79	0.78	0.78
ISUP3 class	0.87	0.77	0.81
ISUP4 class	0.86	0.94	0.89
Average	0.84	0.85	0.84

According to Table 2, the F-scores are high in the following order: ISUP4 class, ISUP1 class, ISUP3 class, and ISUP2 class. If we look back at how many times we augmented the classes, they are high in this order: fourteen times for the ISUP4 class, eleven times for the ISUP1 class, five times for the ISUP3 class, and twice for the ISUP2 class. We can assert that higher accuracy metrics are obtained from a class when there is more augmentation in that class. It arises because the predicted classes for augmented images are frequently the same as those for the original patient image. Most of the time, if the ISUP class of the original image could be accurately recognized, it could also be accurately detected for the augmented version.

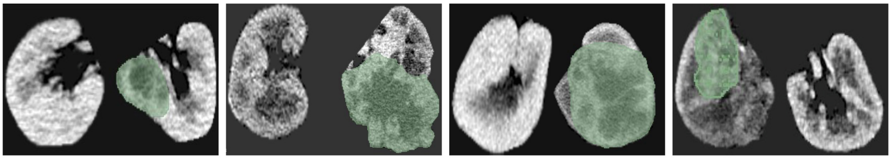
It may be beneficial since it demonstrates how the model can recognize that the augmented image is another version of the original image and forecast the same ISUP class for it. If we look at the accuracy metrics for the ISUP2 class, they are at their lowest, where data augmentation was used the least compared to the other ISUP classes.

Figure 4 illustrates four images from different ISUP classes that are truly classified, and Fig. 5 illustrates four images that are falsely classified. In Fig. 5, we wrote the true ISUP classes as the caption, and the predicted ISUP classes from left to right are ISUP2, ISUP4, ISUP4, and ISUP2. The green parts of the images are the tumor parts. In Fig. 5b, despite the large tumor size, the true ISUP grade was two, and the model identified ISUP 4 in this image. In Fig. 5d, the tumor size was small; the true ISUP class for this image was four, but the model predicted ISUP 2. It demonstrates the model's attempt to concentrate on tumor sizes in its prediction.



(a) ISUP1 class (b) ISUP2 class (c) ISUP3 class (d) ISUP4 class

Fig. 4. Correctly classified images



(a) ISUP1 class (b) ISUP2 class (c) ISUP3 class (d) ISUP4 class

Fig. 5. Misclassified images

Based on a few tumor features, the ISUP grade is determined. When you ask a physician to determine the ISUP class based only on observing CT scan images, they are unable to do so with high certainty [5,18]. We attempted to create a model that could look at patients' CT scans and forecast their ISUP classes. We can conclude that our model was able to extract hidden features relevant to ISPU classes that might not be seen by human eyes.

It is worth motioning that this study has some limitations: 1) to predict ISUP grade, our model needs to get information as input from both the two kidneys and manually segmented tumor(s) indicating the location of the kidney tumors. There is an extract pre-processing stage that extracts the kidneys from the input volume using the manual segmentations of the two kidneys. Our proposed framework might not be able to produce highly accurate classification results from the whole abdominal volumes, and 2) we noticed that sometimes our trained model tries to predict ISUP classes by looking at the tumor size. This impurity leads to ISUP misclassification, so small tumors with grade 4 surface patterns might be classified as grade 1 or 2.

5 Future Work

We can apply transfer learning to improve the performance results [38,39] by getting more three-dimensional images from the cancer imaging archive. Any 3D medical imaging, such as an MRI of the brain or liver, can be used, but for better outcomes, kidney images should be included [40].

Furthermore, we can utilize our image classification layers before fully connected layers as feature extractors since we can use convolutional neural networks for feature extraction [41]. We can link these features to the survival features since the outputs of our image classification are related to the risk score. Thus, we would provide the survival features as the input to the DL-based survival functions, and we can estimate the time of the patient's death by using the patient's medical images.

6 Conclusion

In this study, we proposed a classification framework for kidney tumors based on the International Society of Urological Pathology (ISUP) grading system. We transformed 2D EfficientNet-B7 into a 3D variant that can handle 3D data volumes. To enhance the classification performance, we applied various data augmentation and pre-processing methods. We eliminated other organs in the volumes and kept only the kidneys. The extracted kidneys were concatenated with the provided manual ground-truth annotations of the tumors. This image concatenation is shown to be an important step to force our 3D EfficientNet-B7 to look particularly at the tumors' surface patterns and associate them with the ISUP classes. The data augmentation was applied to first increase the number of samples in the training set and second to partially solve the class imbalance issue. Several image pre-processing methods were applied to enhance the input image quality. The proposed framework demonstrated good classification accuracy of (84%) on the test set. This study shows how crucial it is to properly prepare the dataset through actions like cropping, augmentation, and pre-processing. It is worth mentioning that we tried to show how the results of this work can be generalized to other datasets as well. However, we could not find any similar dataset in which we could get the required information, such as MRI or CT images of the organs, organ segmentation, tumor ground truth, and, most importantly, ISUP grades.

Acknowledgement. This research was funded by the research council Norway and ICT:5G-HEART. We thank our colleagues Davit Aghayan and Egidijus Pelanis from Intervention Center, Oslo University Hospital, who provided insight and expertise in medical images and clinical aspects that greatly assisted the research. We gratitude Piotr Bialecki, Senior Engineering Manager of the PyTorch Team at NVIDIA, for assistance with technical parts of the training model with deep neural networks; and Håvard Kvamme, previous Ph.D. student of the University of Oslo in the faculty of Mathematics, for his comments and showing the actual path of the research.

References

1. Sung, H., et al.: Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J. Clin. **71**(3), 209–249 (2021)
2. Molina, A.M., et al.: A phase 1b clinical trial of the multi-targeted tyrosine kinase inhibitor lenvatinib (e7080) in combination with everolimus for treatment of metastatic renal cell carcinoma (RCC). Cancer Chemother. Pharmacol. **73**(1), 181–189 (2014)
3. Motzer, R.J., et al.: Dovitinib versus sorafenib for third-line targeted treatment of patients with metastatic renal cell carcinoma: an open-label, randomised phase 3 trial. Lancet Oncol. **15**(3), 286–296 (2014)
4. Samaratunga, H., Gianduzzo, T., Delahunt, B.: The ISUP system of staging, grading and classification of renal cell neoplasia. J. Kidney Cancer VHL **1**(3), 26 (2014)
5. Warren, A.Y., Harrison, D.: WHO/ISUP classification, grading and pathological staging of renal cell carcinoma: standards and controversies. World J. Urol. **36**, 1913–1926 (2018)
6. Rees, M., Tekkis, P.P., Welsh, F.K., O'rourke, T., John, T.G.: Evaluation of long-term survival after hepatic resection for metastatic colorectal cancer: a multifactorial model of 929 patients. Ann. Surg. **247**(1), 125–135 (2008)
7. Zhu, M., Ren, B., Richards, R., Suriawinata, M., Tomita, N., Hassanpour, S.: Development and evaluation of a deep neural network for histologic classification of renal cell carcinoma on biopsy and surgical resection slides. Sci. Rep. **11**(1), 1–9 (2021)
8. Abdeltawab, H.A., Khalifa, F.A., Ghazal, M.A., Cheng, L., El-Baz, A.S., Gondim, D.D.: A deep learning framework for automated classification of histopathological kidney whole-slide images. J. Pathol. Inform. **13**, 100093 (2022)
9. Abu Haeyeh, Y., Ghazal, M., El-Baz, A., Talaat, I.M.: Development and evaluation of a novel deep-learning-based framework for the classification of renal histopathology images. Bioengineering **9**(9), 423 (2022)
10. Fenstermaker, M., Tomlins, S.A., Singh, K., Wiens, J., Morgan, T.M.: Development and validation of a deep-learning model to assist with renal cell carcinoma histopathologic interpretation. Urology **144**, 152–157 (2020)
11. Han, S., Hwang, S.I., Lee, H.J.: The classification of renal cancer in 3-phase CT images using a deep learning method. J. Digit. Imaging **32**(4), 638–643 (2019)
12. Xi, I.L., et al.: Deep learning to distinguish benign from malignant renal lesions based on routine MR ImagingDeep learning for characterization of renal lesions. Clin. Cancer Res. **26**(8), 1944–1952 (2020)
13. Baghdadi, A., et al.: Automated differentiation of benign renal oncocytoma and chromophobe renal cell carcinoma on computed tomography using deep learning. BJU Int. **125**(4), 553–560 (2020)
14. Nikpanah, M., et al.: A deep-learning based artificial intelligence (AI) approach for differentiation of clear cell renal cell carcinoma from oncocytoma on multi-phasic MRI. Clin. Imaging **77**, 291–298 (2021)
15. Zhou, L., Zhang, Z., Chen, Y.-C., Zhao, Z.-Y., Yin, X.-D., Jiang, H.-B.: A deep learning-based radiomics model for differentiating benign and malignant renal tumors. Transl. Oncol. **12**(2), 292–300 (2019)
16. Hadjiyski, N.: Kidney cancer staging: deep learning neural network based approach. In: 2020 International Conference on e-Health and Bioengineering (EHB), pp. 1–4. IEEE (2020)

17. Hussain, M.A., Hamarneh, G., Garbi, R.: Renal cell carcinoma staging with learnable image histogram-based deep neural network. In: Suk, H.-I., Liu, M., Yan, P., Lian, C. (eds.) MLMI 2019. LNCS, vol. 11861, pp. 533–540. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32692-0_61
18. Delahunt, B., Cheville, J.C., et al.: The international society of urological pathology (ISUP) grading system for renal cell carcinoma and other prognostic parameters. Am. J. Surg. Pathol. **37**, 1490–1504 (2013)
19. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks (2019)
20. Tian, K., et al.: Automated clear cell renal carcinoma grade classification with prognostic significance. PLoS ONE **14**(10), e0222641 (2019)
21. Yeh, F.-C., Parwani, A.V., Pantanowitz, L., Ho, C.: Automated grading of renal cell carcinoma using whole slide imaging. J. Pathol. Inform. **5**(1), 23 (2014)
22. Sun, X., et al.: Prediction of ISUP grading of clear cell renal cell carcinoma using support vector machine model based on CT images. Medicine **98**(14) (2019)
23. Cui, E., et al.: Predicting the ISUP grade of clear cell renal cell carcinoma with multiparametric MR and multiphase CT radiomics. Eur. Radiol. **30**, 2912–2921 (2020)
24. Zhao, Y., et al.: Deep learning based on MRI for differentiation of low- and high-grade in low-stage renal cell carcinoma. J. Magn. Reson. Imaging **52**(5), 1542–1549 (2020)
25. Heller, N., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the kits19 challenge. Med. Image Anal. 101821 (2020)
26. Heller, N., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes (2019)
27. Zhao, H., Li, H., Cheng, L.: Chapter 14 - data augmentation for medical image analysis. In: Burgos, N., Svoboda, D. (eds.) Biomedical Image Synthesis and Simulation. The MICCAI Society book Series, pp. 279–302. Academic Press (2022)
28. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
29. Shahinfar, S., Meek, P., Falzon, G.: "How many images do i need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. Ecol. Inform. **57**, 101085 (2020)
30. Pérez-García, F., Sparks, R., Ourselin, S.: TorchIO: a python library for efficient loading, pre-processing, augmentation and patch-based sampling of medical images in deep learning. Comput. Methods Programs Biomed. **208**, 106236 (2021)
31. Akar, E., Kara, S., Akdemir, H., Kiriş, A.: Fractal analysis of MR images in patients with chiari malformation: the importance of pre-processing. Biomed. Signal Process. Control **31**, 63–70 (2017)
32. Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K.: Recurrent residual U-Net for medical image segmentation. J. Med. Imaging (Bellingham) **6**, 014006 (2019)
33. Vankdothu, R., Hameed, M.A.: Brain tumor MRI images identification and classification based on the recurrent convolutional neural network. Meas. Sens. 100412 (2022)
34. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
35. Tustison, N.J., et al.: N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging **29**, 1310–1320 (2010)

36. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014)
37. Boone, L., et al.: ROOD-MRI: benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in MRI (2022)
38. Krishna, S.T., Kalluri, H.K.: Deep learning and transfer learning approaches for image classification (2019)
39. Shaha, M., Pawar, M.: Transfer learning for image classification. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 656–660 (2018)
40. Hussain, M., Bird, J.J., Faria, D.R.: A study on CNN transfer learning for image classification. In: Lotfi, A., Bouchachia, H., Gegov, A., Langensiepen, C., McGinnity, M. (eds.) UKCI 2018. AISC, vol. 840, pp. 191–202. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-97982-3_16
41. Yang, A., Yang, X., Wu, W., Liu, H., Zhuansun, Y.: Research on feature extraction of tumor image based on convolutional neural network. IEEE Access **7**, 24204–24213 (2019)

IoT-HR: Internet of Things in Health Research

An IoT-Based System for the Study of Neuropathic Pain in Spinal Cord Injury

Dario Salvi[1](✉), Gent Ymeri[1], Daniel Jimeno[2], Vanesa Soto-León[3],
Yolanda Pérez-Borrego[3], Carl Magnus Olsson[1],
and Carmen Carrasco-Lopez[1]

[1] Internet of Things and People, Malmö University, Malmö, Sweden
`dario.salvi@mau.se`
[2] Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain
[3] FENNSI Group, National Hospital for Paraplegics, Toledo, Spain

Abstract. Neuropathic pain is a difficult condition to treat and would require reliable biomarkers to personalise and optimise treatments. To date, pain levels are mostly measured with subjective scales, but research has shown that electroencephalography (EEG) and heart rate variability (HRV) can be linked to those levels. Internet of Things technology could allow embedding EEG and HRV in easy-to-use systems that patients can use at home in their daily life. We have developed a system for home monitoring that includes a portable EEG device, a tablet application to guide patients through imaginary motor tasks while recording EEG, a wearable HRV sensor and a mobile phone app to report pain levels. We are using this system in a clinical study involving 15 spinal cord injury patients for one month. Preliminary results show that relevant data are being collected, with inter and intra-patients variability for both HRV and pain levels, and that the mobile phone app is perceived as usable, of good quality and useful. However, because of its complexity, the system requires some effort from patients, is sometimes unreliable and the collected EEG signals are not always of the desired quality.

Keywords: IoT · EEG · HRV · Neuropathic pain · mobile health

1 Introduction

Neuropathic Pain (NP) is defined by the International Association for the Study of Pain as "pain caused by a lesion or disease of the somatosensory nervous system" [1] and is considered one of the most difficult painful conditions to treat [3]. NP is present, for example, in Spinal Cord Injury (SCI), with a prevalence ranging between 40% [27] and 60% [32]. Unfortunately, pharmaceutical treatments for NP have limited efficacy (<50%) [8], hence an objective and robust biomarker for NP is desirable to personalise and optimise treatments.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 93–103, 2023.
https://doi.org/10.1007/978-3-031-34586-9_7

The measurement of the intensity of NP is based on psychometric scales and quality of life questionnaires [6,10]. This introduces subjectivity and only a partial/static view of the wide circumstances of the patients, which also affects the estimation of the effectiveness of treatments both in clinics and trials. To avoid bias, daily records of the level of pain become a tool to complement the proper evaluation of the patient.

Given that pain causes autonomic responses, physiological variables can be measured as biomarkers for NP, such as blood pressure, skin conductance, respiration heart rate, and electroencephalography (EEG) [14,18]. Particularly heart rate variability (HRV) shows a clear link with pain [13], due to the decrease in parasympathetic activation, which causes decreased high-frequency HRV [31] and has been even associated with the subjective experience of pain [9]. In addition to HRV, a recent systematic review on biomarkers for chronic NP [19] identified EEG as a candidate method for measuring NP objectively. These findings suggest that the combined use of EEG and HRV could provide a reliable, objective quantification of NP, as proven, for example, in placebo analgesia [7].

Internet of Things technologies could provide a means to measure these quantities with high frequency and at patients' homes. HRV can nowadays be acquired continuously by inexpensive wrist-worn fitness trackers and smartwatches [20], which can share data through their companion smartphone apps. HRV, however, can be affected by several factors in addition to pain [36], and should therefore be completed with EEG, which can be delivered with modern portable EEG devices [37], and used during scheduled measurement sessions guided through a personal computer or smartphone [2].

In terms of mobile and IoT-based systems for pain measurement and management, the literature review shows a scarcity of proven solutions, notwithstanding the growing interest in the matter [21]. Typical applications include electronically delivered surveys and scales, training programs for treatment and self-management, remote consultations, rehabilitation, psychological support and therapies, medication management and adherence [16,29,30]. Similarly, systems using objective sensor data are scarce in extant research. The aim of this study is therefore to develop an IoT-based system for the study of pain in home settings using EEG and HRV as objective measurements of pain.

This paper is structured as follows: Sect. 2 describes previous work, Sect. 3 describes the developed system while Sect. 4 provides preliminary statistics collected from our running clinical study. Finally, Sect. 5 concludes the findings so far and outlines opportunities for future development.

2 Previous Work

Machine learning has been used to identify spinal cord injured participants at risk of developing central NP from multichannel EEG [35]. Three classifiers (artificial neural networks ANN, support vector machine SVM and linear discriminant analysis LDA) were shown to obtain similar results with higher than 85% classification accuracy on a full set of features. Similarly, using a Support Vector

Machine algorithm has been proven to allow differentiation between patients with chronic pain and healthy controls [34]. This showed an accuracy of >85% solely based on the brain activity of three regions of interest: somatosensory cortex and pregenual and dorsal anterior cingulate cortices.

HRV has been studied on electrocardiogram collected from healthy subjects and patients with and without NP at rest [12]. Results show that participants with NP exhibit a lower HRV, as determined by the standard deviation in R-R length. Studies of electronic systems and IoT have used wearable accelerometers to quantify daily activity in patients with pain [4,28,33]. These highlight that there are differences in activity and behaviour between patients and healthy controls, thus outlining the direction as promising.

In order to detect and quantify pain, one approach has been to measure heart rate, skin conductance, skin temperature and respiratory rate [22]. The data were then combined using an Artificial Neural Network (ANN) and a Fuzzy expert system. Additionally, this research employed a mobile phone app to collect data and provide first-hand assistance. Another example of combining multiple sensors to recognise pain level in healthy volunteers subjected to painful heat stimuli include taking signals from 3 cameras, a Kinect, facial electromyography (EMG), skin conductance level and electrocardiogram [11]. These data were used to train machine learning algorithms where ECG was used to extract features from RR intervals, similarly to how HRV is computed.

Finally, facial EMG has also been used to implement a novel sensor based on a flexible printed circuit board which naturally fits a human's face [38]. This system includes a mobile phone app to receive the sensor data through WiFi, and a cloud-based architecture where to store, review and process patients' data in almost-real time. The proposed solution is, however, impractical in the long term.

While the literature review shows a supportive stance on the idea of combining several sensors' data to measure pain, the majority of proposed systems are prototypes tested with healthy individuals subject to induced pain. Very few have been clinically validated with patients with NP and no studies were found where EEG and HRV are combined in those patients.

3 Methods

We have developed an IoT-based system for the collection of HRV, EEG data, and self-assessment of pain level, which is being used in a clinical study with SCI patients.

3.1 Study Protocol

We are recruiting 15 SCI patients for one month each. Subjects are recruited at the National Hospital for Paraplegics, Toledo, Spain. To include a representative and ample sample of the SCI population, the inclusion criteria are age between 18 and 75 years, any aetiology, any level of injury, minimum time post-SCI of 6 weeks, presence of pain for more than 4 weeks, and a pain level between 2 and 8. The exclusion criteria are severe psychiatric disorders, regular drug use, and the impossibility of using the app. This experimental protocol was approved by the ethics committee of the University Hospital Complex of Toledo (No. CEIC-621) and conducted according to the Declaration of Helsinki.

Patients are asked to record their pain level 3 times a day using a mobile-phone-based Visual Analogue Scale (VAS) [5]. The VAS scale is filled in at least 3 times a day (reminders are sent at 8:00, at 14:00, and at 20:00), but patients are also asked to fill in pain levels when they recognise that the pain is increasing, in accordance with a momentary assessment method [23]. Additionally, patients are asked to wear a smartwatch to measure heart rate variability continuously. Patients record their EEG activity once per day (30 days) guided by a tablet application that defines when to rest or perform an imaginary motor task. Fifteen days after starting the study, patients are asked to answer a usability [15] and technology acceptance [17] questionnaire, delivered through the app.

3.2 IoT System

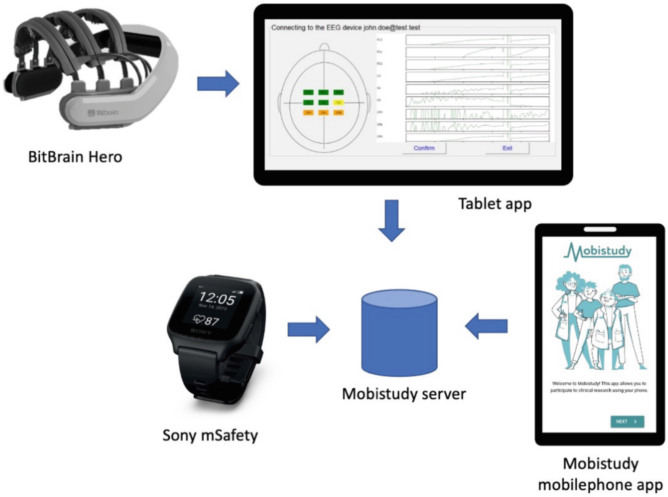

BitBrain Hero
Connecting to the EEG device john.doe@test.test
Confirm
Exit
Tablet app
Mobistudy
Welcome to Mobistudy! This app allows you to participate to clinical research using your phone.
NEXT
12:05
87
SONY
Mobistudy server
Sony mSafety
Mobistudy mobilephone app

Fig. 1. IoT-base system architecture including sensors (Sony mSafety and BitBrain Hero), user applications (Mobistudy app and tablet app) and server.

We developed an IoT system with four modules (see Fig. 1). First, we use the Mobistudy app [25] to allow patients to record their pain levels. Through the app, patients can create a profile and join the PainApp study using an invitation code. The app sends reminders to ask patients to fill in the VAS scale when a task is due, but patients are also able to report their pain level at any time. The

VAS scale was implemented using a horizontal slider with values from 0 to 100. To allow patients to focus on the visual scale rather than the numerical value, the actual chosen value is not shown on the interface. Usability and acceptance questionnaires are also delivered through the app, using its configurable "forms" feature [24].

Second, we use a smartwatch used for collecting HRV data. We use the Sony mSafety wearable device, which is able to measure steps, heart rate, heart rate variability and activity type. We configured the device to measure HRV every 15 min. The data is sent by the watch to data storage using an embedded LTE CAT-M1 radio communication module. Messages are encrypted when sent and unencrypted when downloaded from the mSafety infrastructure.

Third, a tablet application allows patients to perform imaginary motor tasks while measuring brain activity through EEG. As EEG device, we use the wearable and mobile BitBrain Hero EEG headset, with 9 dry electrodes. The tablet application connects to the Hero through a USB or Bluetooth connection, collects the raw EEG data, and guides the user through the tasks defined according to the protocol [26].

Fourth, a backend server collects all data from applications and devices, using the Mobistudy REST API. The mobile phone and tablet applications are integrated with the server, and the data from smartwatches are downloaded from the mSafety infrastructure through a webhook.

4 Preliminary Results and Discussion

The clinical study is still running at the time of writing. As more than half of the patients have completed the study, we present preliminary statistics about usage and information useful to evaluate the reliability and usability of the system, based on the data collected.

Eighteen patients have been involved so far, with 12 having concluded the study and 3 having dropped out for technical reasons: two users could not register EEG data with enough quality and 1 patient did not receive notifications on the phone. Five patients are female and 13 are male. The average age is 46.5 with an 8.7 standard deviation.

Ten patients have contributed with 218 EEG sessions in total. Thirteen patients have contributed with 653 pain level reports using the app and 8 patients have contributed with 5673 HRV measurements since the start of the study.

During the execution of the study, a bug was identified in the Mobistudy app that prevented scheduled notifications to be sent after the first 3 days of participation. The bug has not been completely fixed yet, but, 3 months after the beginning of the study, patients were made aware of it and instructed about how to temporarily circumvent it. This issue had an impact on the number of responses received (which is difficult to quantify) and led to one patient ceasing to report pain levels completely.

A box plot of the received VAS pain levels recorded through the app is shown in Fig. 2. It can be observed that pain levels vary considerably among patients (inter-patient variation) with some patients having wider variations than others (intra-patient variation). The inter-patient variation is promising for associating EEG and HRV features with pain levels as greater variability will facilitate the development and validation of algorithms able to associate features on a wider scale. These results suggest that patients with a high level of pain are indicative of ineffective treatment and that intra-patient variation can indicate the need for a more adjusted treatment regime, for example, to contrast the effect of the medication gradually fading during the day. This exemplifies the potential usefulness of telemonitoring systems in clinical practice for NP.

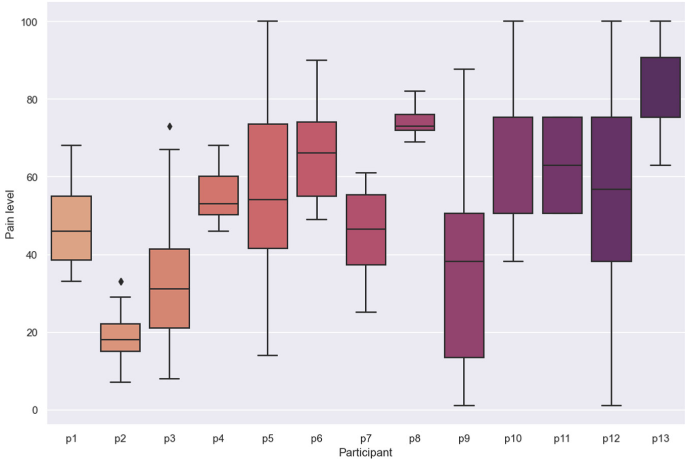


Fig. 2. Distribution (box plots) of the pain levels as reported on the mobile phone app for each patient.

The distribution of HRV measurements is shown in Fig. 3 for those 8 patients who have provided the data so far. Here, the main observation is that most values are within reasonable limits (between 20 and 80 ms) and both inter and intra-patients variability are present. No clear relationship can be derived from the charts between HRV and pain levels, therefore a more detailed analysis will be required in the future, for example by extracting low and high-frequency components of the HRV signals [13,31].

In terms of EEG sessions, we found that maintaining the connection between the tablet, the portable EEG device and the Mobistudy server was not sufficiently reliable. For this reason, we decided to avoid sending the data to the server when patients were recruited and opted to extract the EEG files from the tablet when returned.

Aside from some technical issues such as this, most patients have been able to record their EEG using this new technology. However, patients with a higher degree of motor impairment required assistance in the use of the system. Thus

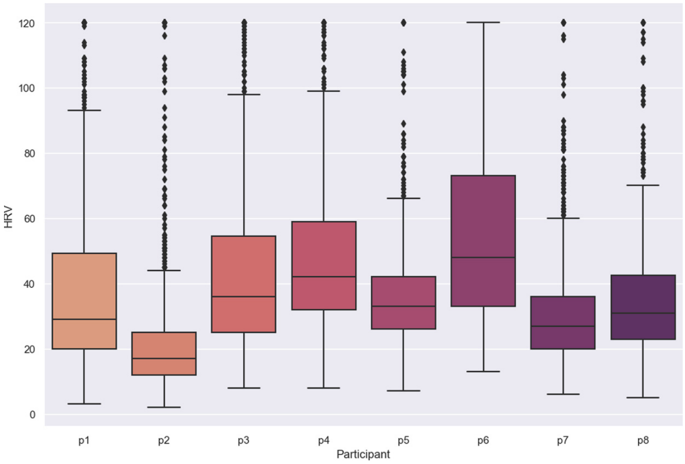


Fig. 3. Distribution (box plots) of the heart rate variability (in ms) measured by the wearable device, per patient.

far, four patients have brought the tablet and the EEG device at home while others were in-patients and therefore helped by a lab technician during the recordings if needed.

EEG sessions are currently being analysed in greater depth by discarding sections of the signal corrupted by noise and by computing spectral density characteristics that will be fed into machine learning algorithms. This is done as we are noticing that several sections of the collected EEG signals have spurious frequencies which we believe are due to poor cabling. This is supported by having used both models of the Hero device with USB and Bluetooth connections, and noticing that the wireless model produces cleaner recordings. Of the 198 sessions from 9 patients analysed so far, 91 sessions (46%) have been discarded because of bad signal quality.

Results from the usability and technology acceptance questionnaires are provided in Tables 1 and 2 respectively. Descriptive statistics about the answers provided to each section show that the app was well received among patients. In terms of quality (uMARS questionnaire [15]), the app was considered engaging, and functional, with a good degree of information, appealing aesthetics and of good quality. Even if positive, engagement scored the lowest among the categories, which shows that filling in the VAS scale 3 times per day may be perceived as tedious.

In terms of long-term acceptance (MOHTAM questionnaire [17]), both usability and perceived usefulness were evaluated high, which indicates that, notwithstanding the repetitiveness of the task, patients find it useful and would be happy to perform it in the long term. This is also confirmed by the answers related to the perceived impact of the uMARS questionnaire, as patients perceive that the app facilitates both the reporting of their pain levels and the participation in clinical studies.

Table 1. Mean and standard deviation of the answers provided to the uMARS questionnaire, by category, on a Likert scale from 1 (negative evaluation) to 5 (positive evaluation). N of patients = 7.

Category	Mean	Standard deviation
Engagement (q1–q5)	3.33	1.22
Functionality (q6–q9)	4.56	0.58
Aesthetics (q10–q12)	4.33	0.58
Information (q13–q16)	4.40	0.71
Subjective quality (q17–q20)	3.39	1.42
Perceived impact in monitoring pain levels (q21–q26)	3.37	1.33
Perceived impact in participating in clinical studies (q27–q32)	4.07	1.45

Table 2. Mean and standard deviation of the answers provided to the technology acceptance questionnaire, by category, on a Likert scale from 1 (negative evaluation) to 5 (positive evaluation). N of patients = 7.

Category	Mean	Standard deviation
Perceived ease of use (q1–q7)	3.55	0.98
Perceived usefulness (q8–q12)	3.31	0.79

5 Conclusions and Future Work

This work presents an IoT-based system for recording EEG and HRV as biomarkers for pain levels in spinal cord injury, used as part of a clinical study. The system uses a complex setup consisting of a portable EEG device, a smartwatch, a tablet application and a mobile phone application. So far, fifteen patients have been using the system successfully, showing that data can be collected and that the app is usable. While collection and analysis of the data is ongoing work, our preliminary results show that the data can be used to profile patients in a clinically meaningful way.

Nonetheless, the complexity of setting up the devices, connections, and overall usability of this novel technology is still a challenge that needs to be addressed. Over time, such work needs to ensure ease of use for the intended patient group as well as to improve and validate the reliability of devices included in the system to minimize the risk of corrupted signals.

Acknowledgements. The work presented in this paper has been partially funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 101030384, by the Swedish Knowledge Foundation through the Internet of Things and People research center, and thanks to sponsorship from Bitbrain and

Sony. We would like to thank Rohan Samandari for his contributions to the development of the tablet application.

References

1. Abrecht, C.R., Nedeljkovic, S.S.: Neuropathic pain. In: Yong, R., Nguyen, M., Nelson, E., Urman, R. (eds.) Pain Medicine, pp. 541–543. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-43133-8_145
2. Anwar, D., Garg, P., Naik, V., Gupta, A., Kumar, A.: Use of portable EEG sensors to detect meditation. In: 2018 10th International Conference on Communication Systems & Networks (COMSNETS), pp. 705–710. IEEE (2018)
3. Attal, N., Bouhassira, D., Baron, R.: Diagnosis and assessment of neuropathic pain through questionnaires. Lancet Neurol. **17**(5), 456–466 (2018)
4. van den Berg-Emons, R.J., Schasfoort, F.C., de Vos, L.A., Bussmann, J.B., Stam, H.J.: Impact of chronic pain on everyday physical activity. Eur. J. Pain **11**(5), 587–593 (2007)
5. Crichton, N.: Visual analogue scale (VAS). J. Clin. Nurs. **10**(5), 706–6 (2001)
6. Cruccu, G., et al.: EFNS guidelines on neuropathic pain assessment. Eur. J. Neurol. **11**(3), 153–162 (2004)
7. De Pascalis, V., Vecchio, A.: The influence of EEG oscillations, heart rate variability changes, and personality on self-pain and empathy for pain under placebo analgesia. Sci. Rep. **12**(1), 1–18 (2022)
8. Finnerup, N.B., et al.: Pharmacotherapy for neuropathic pain in adults: a systematic review and meta-analysis. Lancet Neurol. **14**(2), 162–173 (2015)
9. Forte, G., Troisi, G., Pazzaglia, M., Pascalis, V.D., Casagrande, M.: Heart rate variability and pain: a systematic review. Brain Sci. **12**(2), 153 (2022)
10. Haanpää, M., et al.: Neupsig guidelines on neuropathic pain assessment. PAIN® **152**(1), 14–27 (2011)
11. Kächele, M., Werner, P., Al-Hamadi, A., Palm, G., Walter, S., Schwenker, F.: Bio-visual fusion for person-independent recognition of pain intensity. In: Schwenker, F., Roli, F., Kittler, J. (eds.) MCS 2015. LNCS, vol. 9132, pp. 220–230. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20248-8_19
12. Karri, J., Zhang, L., Li, S., Chen, Y.T., Stampas, A., Li, S.: Heart rate variability: a novel modality for diagnosing neuropathic pain after spinal cord injury. Front. Physiol. **8**, 495 (2017)
13. Koenig, J., Jarczok, M., Ellis, R., Hillecke, T., Thayer, J.F.: Heart rate variability and experimentally induced pain in healthy adults: a systematic review. Eur. J. Pain **18**(3), 301–314 (2014)
14. Loggia, M.L., Juneau, M., Bushnell, M.C.: Autonomic responses to heat pain: heart rate, skin conductance, and their relation to verbal ratings and stimulus intensity. PAIN® **152**(3), 592–598 (2011)
15. Martin-Payo, R., Carrasco-Santos, S., Cuesta, M., Stoyan, S., Gonzalez-Mendez, X., Fernandez-Alvarez, M.D.M.: Spanish adaptation and validation of the user version of the mobile application rating scale (uMARS). J. Am. Med. Inform. Assoc. **28**(12), 2681–2686 (2021)
16. McGeary, D.D., McGeary, C.A., Gatchel, R.J.: A comprehensive review of telehealth for pain management: where we are and the way ahead. Pain Pract. **12**(7), 570–577 (2012)

17. Mohamed, A.H.H., Tawfik, H., Al-Jumeily, D., Norton, L.: MoHTAM: a technology acceptance model for mobile health applications. In: 2011 Developments in E-systems Engineering, pp. 13–18. IEEE (2011)
18. Möltner, A., Hölzl, R., Strian, F.: Heart rate changes as an autonomic component of the pain response. Pain **43**(1), 81–89 (1990)
19. Mussigmann, T., Bardel, B., Lefaucheur, J.P.: Resting-state electroencephalography (EEG) biomarkers of chronic neuropathic pain. a systematic review. NeuroImage 119351 (2022)
20. Natarajan, A., Pantelopoulos, A., Emir-Farinas, H., Natarajan, P.: Heart rate variability with photoplethysmography in 8 million individuals: a cross-sectional study. Lancet Digit. Health **2**(12), e650–e657 (2020)
21. Prada, E.J.A.: The internet of things (IoT) in pain assessment and management: an overview. Inform. Med. Unlock. **18**, 100298 (2020)
22. Rajesh, M., Muthu, J.S., Suseela, G.: iPainRelief-a pain assessment and management app for a smart phone implementing sensors and soft computing tools. In: 2013 International Conference on Information Communication and Embedded Systems (ICICES), pp. 434–441. IEEE (2013)
23. Rost, S., Van Ryckeghem, D.M., Koval, P., Sütterlin, S., Vögele, C., Crombez, G.: Affective instability in patients with chronic pain: a diary approach. Pain **157**(8), 1783–1790 (2016)
24. Salvi, D., Lee, J., Velardo, C., Goburdhun, R.A., Tarassenko, L.: Mobistudy: an open mobile-health platform for clinical research. In: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 918–921. IEEE (2019)
25. Salvi, D., Olsson, C.M., Ymeri, G., Carrasco-Lopez, C., Tsang, K.C., Shah, S.A.: Mobistudy: mobile-based, platform-independent, multi-dimensional data collection for clinical studies. In: 11th International Conference on the Internet of Things, pp. 219–222 (2021)
26. Samandari, R.: Integration of bluetooth sensors in a windows-based research platform. Bachelor's thesis, Malmö University (2021)
27. Siddall, P.J., McClelland, J.M., Rutkowski, S.B., Cousins, M.J.: A longitudinal study of the prevalence and characteristics of pain in the first 5 years following spinal cord injury. Pain **103**(3), 249–257 (2003)
28. Spenkelink, C., Hutten, M.M., Hermens, H., Greitemann, B.O.: Assessment of activities of daily living with an ambulatory monitoring system: a comparative study in patients with chronic low back pain and nonsymptomatic controls. Clin. Rehabil. **16**(1), 16–26 (2002)
29. Sundararaman, L.V., Edwards, R.R., Ross, E.L., Jamison, R.N.: Integration of mobile health technology in the treatment of chronic pain: a critical review. Regional Anesth. Pain Med. **42**(4), 488–498 (2017)
30. Thurnheer, S.E., Gravestock, I., Pichierri, G., Steurer, J., Burgstaller, J.M.: Benefits of mobile apps in pain management: systematic review. JMIR Mhealth Uhealth **6**(10), e11231 (2018)
31. Tracy, L.M., Ioannou, L., Baker, K.S., Gibson, S.J., Georgiou-Karistianis, N., Giummarra, M.J.: Meta-analytic evidence for decreased heart rate variability in chronic pain implicating parasympathetic nervous system dysregulation. Pain **157**(1), 7–29 (2016)
32. Van Gorp, S., Kessels, A., Joosten, E., Van Kleef, M., Patijn, J.: Pain prevalence and its determinants after spinal cord injury: a systematic review. Eur. J. Pain **19**(1), 5–14 (2015)

33. Van Weering, M., Vollenbroek-Hutten, M., Tönis, T., Hermens, H.: Daily physical activities in chronic lower back pain patients assessed with accelerometry. Eur. J. Pain **13**(6), 649–654 (2009)
34. Vanneste, S., De Ridder, D.: Chronic pain as a brain imbalance between pain input and pain suppression. Brain Commun. **3**(1), fcab014 (2021)
35. Vuckovic, A., Gallardo, V.J.F., Jarjees, M., Fraser, M., Purcell, M.: Prediction of central neuropathic pain in spinal cord injury based on EEG classifier. Clin. Neurophysiol. **129**(8), 1605–1617 (2018)
36. Xhyheri, B., Manfrini, O., Mazzolini, M., Pizzi, C., Bugiardini, R.: Heart rate variability today. Prog. Cardiovasc. Dis. **55**(3), 321–331 (2012)
37. Xu, J., Zhong, B.: Review on portable EEG technology in educational research. Comput. Hum. Behav. **81**, 340–349 (2018)
38. Yang, G., et al.: IoT-based remote pain monitoring system: from device to cloud platform. IEEE J. Biomed. Health Inform. **22**(6), 1711–1719 (2017)

IoT Smart Shoe Solution for Neuromuscular Disease Monitoring

Davide La Rosa[1], Filippo Palumbo[1](✉), Alessandra Ronca[2], Francesco Sansone[3], Mario Tesconi[4], Alessandro Tonacci[3], and Raffaele Conte[5]

[1] Institute of Information Science and Technologies, National Council of Research (ISTI-CNR), Pisa, Italy
filippo.palumbo@isti.cnr.it
[2] Department of Information Engineering, University of Pisa, Pisa, Italy
[3] Institute of Clinical Physiology, National Research Council of Italy (IFC-CNR), Pisa, Italy
[4] Adatec S.r.l., Navacchio, Pisa, Italy
[5] National Research Council of Italy (CNR), Rome, Italy

Abstract. Recent advances in sensing, processing, and learning of physiological parameters, make the development of non-invasive health monitoring systems increasingly effective, especially in those situations that need particular attention to the usability of devices and software solutions due to the frailty of the target population. In this context, we developed a sensorized shoe that detects significant features in subjects' gait and monitors variations related to an intervention protocol in people affected by Neuromuscular Disorders (NMDs).

This paper outlines the challenges in the field and summarizes the approach used to overcome the technological barriers related to connectivity, deployment, and usability that are typical in a medical setting. The proposed solution adopts the new paradigm offered by Web Bluetooth based on Bluetooth WebSocket.

We show the architectural and deployment choices and how this solution can be easily adapted to different devices and scenarios.

Keywords: Web Bluetooth · Smart Shoe · IoT Health Device

1 Introduction

Neuromuscular Disorders (NMDs) include a wide range of health conditions affecting the function of muscular structures. They are related to the changes in the muscle or in the peripheral nerves sending signals to the muscles. Their diagnosis and clinical representation are often difficult even for skilled, experienced physicians. This is mainly due to the hundreds of specific NMDs present in the clinical experience, which are classified according to various principles. One of the most widely accepted ones distinguishes between: i) muscular dystrophies; ii) myopathies; iii) neuromuscular junction disorders; iv) motor/sensory neuropathies [1]. However, clinical hallmarks of such disorders are often similar, making their differential diagnosis troublesome. Also, the development of non-intrusive methods for their diagnosis and monitoring is an open challenge, making

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 104–115, 2023.
https://doi.org/10.1007/978-3-031-34586-9_8

the need for easy-to-use and affordable devices and evaluation tests a key aspect of such a scenario.

To this end, we tried to merge the outcomes of two main projects, namely InGene 2.0 and Ki-Foot, both from the clinical and sensing perspective to offer a technological solution that, using a set of a few state-of-the-art short exercises wearing a sensorized shoe, can detect and monitor the evolution of one of the most widely studied conditions in people affected by NMDs, the Facioscapulohumeral muscular dystrophy (FSHD).

From a technological point of view, the shoe can detect significant features from the gait of the user primarily related to the FSHD condition. In this medical context, one of the main barriers to the adoption of technological solutions like this is the lack of connectivity and the reliability of the data collection. To this end, we developed a novel paradigm offered by the Bluetooth standard, namely Web Bluetooth[1], that allows us to connect to the smart shoe (or the data collection Bluetooth device) directly from the browser without any configuration by the medical personnel that can exploit the functionalities of the smart shoe in real-time while performing the exercises.

The chosen protocol is a specification for Bluetooth APIs to allow websites to communicate with devices in a secure and privacy-preserving way, it is still in a beta version and not completely adopted by all the browsers in their current versions, but its promising capabilities are worth the investigation in the eHealth context. For this reason, we show the chosen architectural and deployment solution in order to give a reference development guide to those interested in the implementation of this paradigm in their monitoring solutions.

1.1 The InGene 2.0 Project

Several attempts have been made to relate the clinical diagnosis of an individual, or a group of subjects, with NMD, to their genotype (i.e., their genetic background), to retrieve similarities, clustering, and differences, which might be a useful add-on to the current clinical practice. This is the main basis for the InGene 2.0 project, funded by Tuscany Region, Italy, under the Bando Salute 2018 call for grants, attempting at making use of technological (both hardware and software) tools to support the clinician in the diagnosis of NMD and the relationship retrieval between genotype and phenotype. The project, involving four clinical centers and two research institutions in the Region, aims at proposing this new paradigm of diagnosis and treatment, fruitfully supported by technology, to the regional and national decision-making, with likely positive outcomes in terms of a correct diagnosis and with a truly person-tailored treatment in such disorders.

Considering the most widely studied conditions in this field, Facioscapulohumeral muscular dystrophy (FSHD) attracts a lot of attention from clinicians for its relative incidence with respect to other NMDs, and for their clinical characteristics. In fact, FSHD is a disorder characterized by muscle weakness and wasting (atrophy). The disorder gets its name from muscles that are affected, namely those in the face (*facio*), around the shoulder blades (*scapulo*), and in the upper arms (humeral)[2]. What is particularly intriguing in FSHD is the presence of a peculiar muscular involvement, represented by

---

[1] https://www.w3.org/community/web-bluetooth/.

[2] https://rarediseases.org/rare-diseases/facioscapulohumeral-muscular-dystrophy/.

a relative weakness of the anterior muscles of the leg, mainly the tibialis anterior, which is of relevance according to the clinicians, also due to their somewhat relationship with the relative clinical severity of the disorder [2, 3]. Sometimes, such a muscular structure reflects minor changes related to the disease course even years before the onset of clear, related clinical signs, making it a useful target for tailored investigations. However, although efforts have been made to develop proper methods for the analysis of muscular involvement, which are usable, informative, and affordable, most studies still rely on the use of Magnetic Resonance Imaging (MRI) tools, which are expensive, somewhat obtrusive and not always well accepted by the patients [4, 5].

To this extent, new methodologies combining a fast, user-compliant, cost-affordable approach to study the tibialis anterior involvement in FSHD are desirable, making one of them the core of the investigation presented here. In particular, we chose a set of specific short exercises performed while wearing the smart shoes in order to detect and monitor such a condition: six minute walk test, ten meters walk, timed up and go, and four steps climbing.

1.2 A Smart Shoe for Gait Analysis

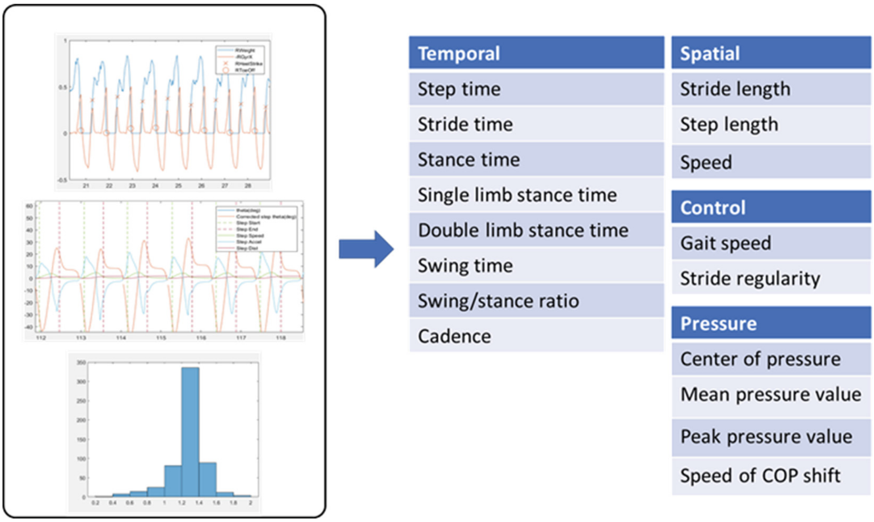
The core sensing device of the overall system is the smart shoes, which are sensorized footwear. Even if smart shoes, from an aesthetic point of view, they are not different from a normal pair of shoes, the upper side is made of leather while the sole belongs to the "Gommus" line, which is a rubber sole line for high performances, high quality, and high design products[3]. Inside the sole, different sensors and communication components[4] are present that allow an accurate analysis of different gait parameters. The Ki-Foot [6] shoe, based on the Motus prototype developed Carlos s.r.l., Fucecchio, Italy, has five pressure sensors integrated under the insole to monitor the mechanical interaction of the foot with the ground: three sensors under the forefoot, and the remaining two under the heel. In this way, almost complete coverage of the entire surface of the sole of the foot is ensured. The pressure sensors are custom-made piezo-resistive transducers produced by using a conductive material on a flexible substrate. These force sensors are sampled at 50 Hz. This kind of sensor has already been tested in different scenarios like sleep monitoring with smart bed slats [7]. A digital Inertial Measurement Unit (IMU) is integrated into the frontal part of the shoe. It consists of a 9-axis inertial platform with a 3D accelerometer, a 3D gyroscope, and a 3D magnetometer. Each value of the accelerometer represents the measure of acceleration of the corresponding axis, and it is measured in mg (milli-gravity). The gyroscope measures the angular speed for each corresponding axis and is expressed in dps (degrees per second). The magnetometer indicates the measurement of the earth's magnetic field for each axis and is expressed in mG (milli-Gauss). A Bluetooth Low Energy (BLE) transmission module is integrated with the rest of the electronic unit in the heel of the shoe to enable low-energy data transmission to a BLE device. The rechargeable battery type LIPO allows the complete operation of the system for 48 h. Being completely wireless controlled, the subject is not conditioned during movement and can move freely and independently for several days. Thanks to these sensors and to

---

[3] http://www.gommus.it.

[4] http://www.adatec.it.

these measures, there is a realistic analysis of the step and the characteristics of the two feet independently. We extract different features (Fig. 1) from the raw data collected to be used by clinicians in their gait analysis while performing the prescribed exercises. In the *Temporal* space, we extract step time, stride time, stance time, single limb stance time, double-limb stance time, swing time, swing-stance ratio, and cadence. The *Spatial* characteristics extracted are stride length, step length, and speed. The *Control* features are gait speed and stride regularity, while in the *Pressure* category falls the parameters related to Center of Pressure (CoP), mean pressure value, peak pressure value, and speed of CoP shift.



Temporal
Step time
Stride time
Stance time
Single limb stance time
Double limb stance time
Swing time
Swing/stance ratio
Cadence

Spatial
Stride length
Step length
Speed

Control
Gait speed
Stride regularity

Pressure
Center of pressure
Mean pressure value
Peak pressure value
Speed of COP shift

Fig. 1. From raw data to gait's temporal, spatial, control, and pressure related features

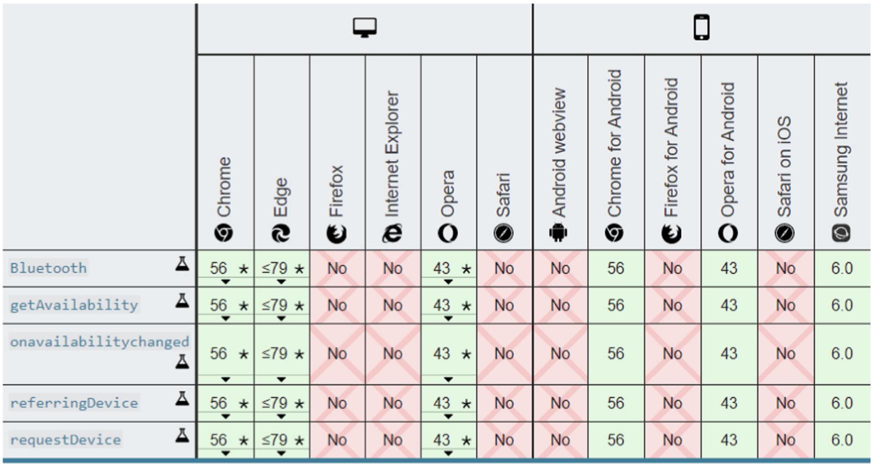
2 The Web Bluetooth Solution

Nowadays, browsers are evolving, bringing new APIs and ways to connect to other devices and allowing access to more functionality than they ever did before. One such API is the Web Bluetooth API[5]. This Web Bluetooth API is still in beta as of this writing, but once this gets released to the public, it will open a whole lot of opportunities for researchers and developers who want to use Bluetooth but don't have the possibility to create a native application for each platform.

The Web Bluetooth API is a low-level API allowing Web applications to pair with the nearby Bluetooth Low Energy-enabled peripheral devices and access their services exposed. Subsets of the Web Bluetooth API are available in some browsers as in Fig. 2. This means it is possible to request and connect to nearby Bluetooth Low Energy devices, read/write Bluetooth characteristics, receive GATT Notifications, know when a Bluetooth device gets disconnected, and even read and write to Bluetooth descriptors.

---

[5] https://www.chromestatus.com/feature/5264933985976320.



	🖥						📱					
	Chrome	Edge	Firefox	Internet Explorer	Opera	Safari	Android webview	Chrome for Android	Firefox for Android	Opera for Android	Safari on iOS	Samsung Internet
`Bluetooth` ⚗	56 ★	≤79 ★	No	No	43 ★	No	No	56	No	43	No	6.0
`getAvailability` ⚗	56 ★	≤79 ★	No	No	43 ★	No	No	56	No	43	No	6.0
`onavailabilitychanged` ⚗	56 ★	≤79 ★	No	No	43 ★	No	No	56	No	43	No	6.0
`referringDevice` ⚗	56 ★	≤79 ★	No	No	43 ★	No	No	56	No	43	No	6.0
`requestDevice` ⚗	56 ★	≤79 ★	No	No	43 ★	No	No	56	No	43	No	6.0

Fig. 2. Compatibility table of the Web Bluetooth functionalities on various browsers and platforms

The *Generic Attribute Profile* (GATT) establishes in detail how to exchange all profile and user data over a BLE connection. In contrast with *Generic Access Profile* (GAP), which defines the low-level interactions with devices, GATT deals only with actual data transfer procedures and formats. GATT also provides the reference framework for all GATT-based profiles, which cover precise use cases and ensure interoperability between devices from different vendors. All standard BLE profiles are therefore based on GATT and must comply with it to operate correctly. This makes GATT a key section of the BLE specification because every single item of data relevant to applications and users must be formatted, packed, and sent according to its rules. GATT uses the *Attribute Protocol* as its transport protocol to exchange data between devices. This data is organized hierarchically in sections called services, which group conceptually related pieces of user data called *Characteristics*.

The GATT *Profile Hierarchy* describes how a GATT *Server* contains a hierarchy of *Profiles*, *Primary Services*, *Included Services*, *Characteristics*, and *Descriptors*.

Profiles are purely logical: the specification of a *Profile* describes the expected interactions between the other GATT entities the *Profile* contains, but it's impossible to query which *Profiles* a device supports.

GATT *Clients* can discover and interact with the *Services*, *Characteristics*, and *Descriptors* on a device using a set of GATT procedures. The specification refers to *Services*, *Characteristics*, and *Descriptors* collectively as *Attributes*. All *Attributes* have a type that's identified by a UUID. Each *Attribute* also has a 16-bit *Attribute Handle* that distinguishes it from other *Attributes* of the same type on the same GATT *Server*. *Attributes* are notionally ordered within their GATT *Server* by their *Attribute Handle*, but while platform interfaces provide attributes in some order, they do not guarantee that it's consistent with the *Attribute Handle* order.

A *Service* contains a collection of *Included Services* and *Characteristics*. The *Included Services* are references to other *Services*, and a single *Service* can be included by more than one other *Service*. *Services* are known as *Primary Services* if they appear directly under the GATT *Server*, and *Secondary Services* if they're only included by other *Services*, but *Primary Services* can also be included. A *Characteristic* contains a value, which is an array of bytes, and a collection of *Descriptors*. Depending on the properties of the *Characteristic*, a GATT *Client* can read or write its value, or register to be notified when the value changes. Finally, a *Descriptor* contains a value (again an array of bytes) that describes or configures its *Characteristic*.

As with any other protocol or profile in the Bluetooth specification, GATT starts by defining the roles that interacting devices can adopt: *Client* or *Server*. The GATT *Client* corresponds to the ATT client using *Attribute Protocol*. It sends requests to a server and receives responses (and server-initiated updates) from it. The GATT client does not know anything in advance about the server's attributes, so it must first inquire about the presence and nature of those attributes by performing service discovery. After completing service discovery, it can then start reading and writing attributes found in the server, as well as receiving server-initiated updates. The GATT *Server* corresponds to the ATT server, which uses *Attribute Protocol* for the connection. It receives requests from a client and sends responses back. It also sends server-initiated updates when configured to do so, and it is the role responsible for storing and making the user data available to the client, organized in attributes. Every BLE device sold must include at least a basic GATT server that can respond to client requests, even if only to return an error response.

The Web Bluetooth API is exposed in the most updated browsers as a Javascript API:

`navigator.bluetooth.requestDevice(serviceFilters)`

Scans for the device in range supporting the requested services. Returns a Promise.

`device.gatt.connect()`

Returns a Promise resolved with the server object providing access to the services available on the device.

`server.getPrimaryService(name)`

Returns a Promise resolved with the particular Bluetooth service on the device.

`service.getCharacteristic(name)`

Returns a Promise resolved with the GATT characteristic object.

`characteristic.readValue()`

Returns a Promise resolved with a raw value from the GATT characteristic.

`characteristic.writeValue(value)`

Writes a new value for the GATT characteristic.

2.1 Web Bluetooth on the Field

Web Bluetooth is not yet a W3C standard but, besides the available implementations in Chrome platforms, roadmaps are available in Mozilla Firefox, Microsoft Edge, and WebKit (Safari)[6]. The API has gained attention in literature, and it has been implemented in various contexts. In [8], a push notification-based login method has been proposed, while in [9], authors present a method for rapid development of applications in distributed BLE IoT systems for eHealth and sports. In that work, a throughput comparison between a native and a Web Bluetooth solution has been presented and the conclusion was that,

---

[6] https://www.bluetooth.com/blog/the-web-bluetooth-series/.

despite the fact that at a higher transmission rate a native application outperforms the HTML5 Web Bluetooth application, the developed Web Bluetooth framework enables software and service operators to iteratively create, tune and deploy filter algorithms in distributed BLE IoT systems, without rebooting nodes or restarting programs using dynamic software updating. Also, a human centered Web-based dataset creation and annotation tool for real time motion detection has been developed [10] in which the user can effectively collect gestures from a nearby device that supports the BLE protocol, assign tags to the collected data, and store them remotely. A particular example of Web Bluetooth implementation in eHealth applications is presented in [11] that present the opportunities for Brain Computer Interaction (BCI) developers to stream data directly to a web browser.

3 The Proposed IoT Solution

Designing an IoT solution that can be used in clinical settings requires a careful evaluation of the technological aspects involved in such a context in terms of connection availability, easiness of deployment and maintenance, and usability by the medical staff or the caregivers. To analyze the possible scenarios and their implications, we considered three main entities that come into play:

- Web App: the application running within a browser on a mobile device, used by the clinicians to record the patients' data as well as handling the execution of the exercises proposed to the users
- BLE shoes: sensorized devices worn by the patients, able both to collect inertial and pressure readings and to transmit the data to a receiving device via a Bluetooth Low-Energy connection
- Logger: service in charge of recording the raw data generated by the sensorized shoes during an exercise and associating the recorded data to a specific user identifier for the subsequent processing phases

Through the Web Bluetooth API, any web application can interact with nearby Bluetooth devices in a secure and privacy-preserving way, without the need to deploy additional platform-specific apps. Depending on the environmental conditions and the features of the involved devices, we identified two reference scenarios in which the designed architecture can provide a feasible and effective solution.

3.1 Scenario 1: Web App as a Proxy

In this scenario, the Web App runs in a browser and acts as a bridge between the Bluetooth shoes and the remote logger service (Fig. 3). The shoes are directly paired with the mobile device and the Web App exploits the Web Bluetooth API provided by the browser to be able to connect to the shoes and receive the generated data in real-time. Subsequently, by employing a web socket connection to a backend device, the app sends the raw data collected from the patient exercise to the remote logger service.

The main interactions among the three involved entities are shown in the sequence diagram of Fig. 4. The Web App, as soon as the clinician has selected the patient and

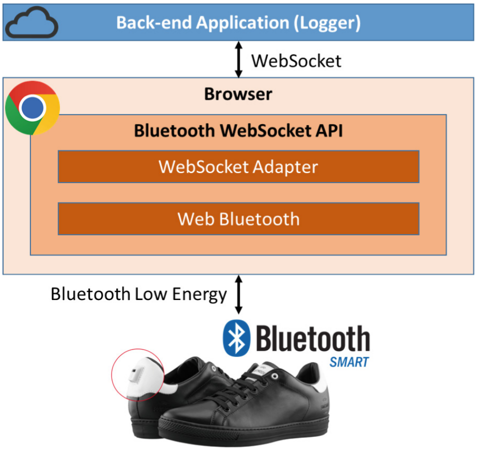


Fig. 3. The Web App collects data from BLE shoes and sends it to the remote logger

started the exercise, requests the available devices to the browser, which in turn triggers a discovery process to find the Bluetooth shoes by name. Whenever the shoes are detected, the Web App queries for both the available services and characteristics. Once the characteristic containing the sensors data is matched, the associated notification is enabled and, since then, the Web App starts to receive and locally store the data stream acquired from the shoes. This loop continues until the clinician stops the exercise from the Web App; then the connection to the shoes is closed and the locally cached recorded exercise is transferred to the remote logger service.

In this scenario the logger can be deployed on a remote backend, thus keeping the number of devices that needs to be deployed on-site at a minimum. This aspect is very important when the technical staff supporting the clinicians could not provide continuous or immediate assistance during the operational phase. On the contrary, we should note that the Bluetooth connection between the sensing devices and the smartphones or tablets can be impaired by potential technical limitations. In our case, adopting a pair of shoes transmitting at a data rate of 50 Hz each in environments where other transmitting devices are present, reduces the reliability of the communication, causing irregular data transmission and occasional connection losses. To improve the situation, we rearranged the system architecture to increase the performance of the Bluetooth data transmission, hence conceiving scenario 2 illustrated in the next section.

3.2 Scenario 2: Single-Board PC as a BLE Device

To overcome the technical limitations of the specific devices we used in our experimental settings, we decided to use two receiving Bluetooth antennas, one for each shoe, to collect data. We used two antennas on a small single-board PC, like for instance the Raspberry Pi, able to host the logger service as well. In this scenario, shown in Fig. 5, the shoes are directly paired with the logger device and the Web App acts as a controller to start and stop the data recording task. The logger device exposes itself as a Bluetooth device to be detected by the browser via the Web Bluetooth API.

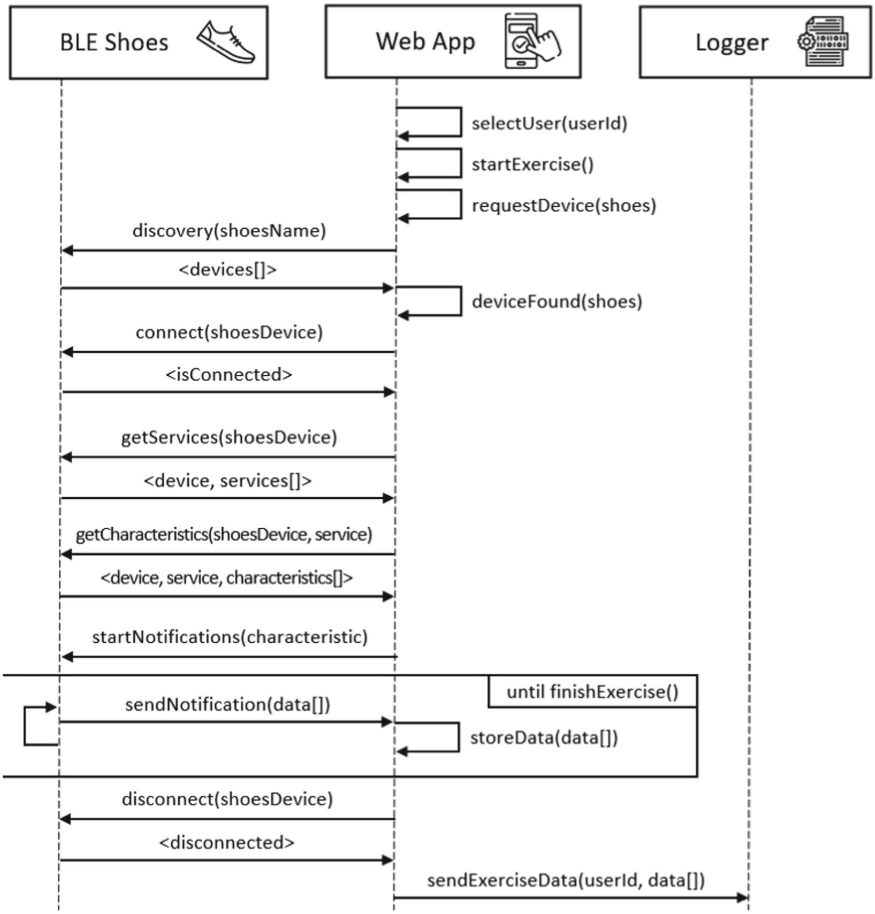


Fig. 4. Sequence diagram of the interaction among the devices in scenario 1

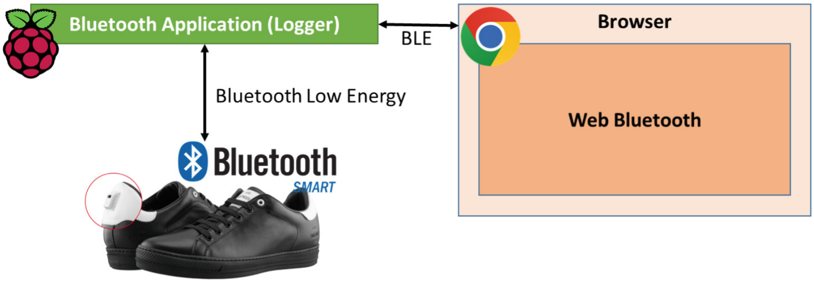


Fig. 5. The Web App connects to the logger service which, in turn, collects the data directly from the BLE shoes

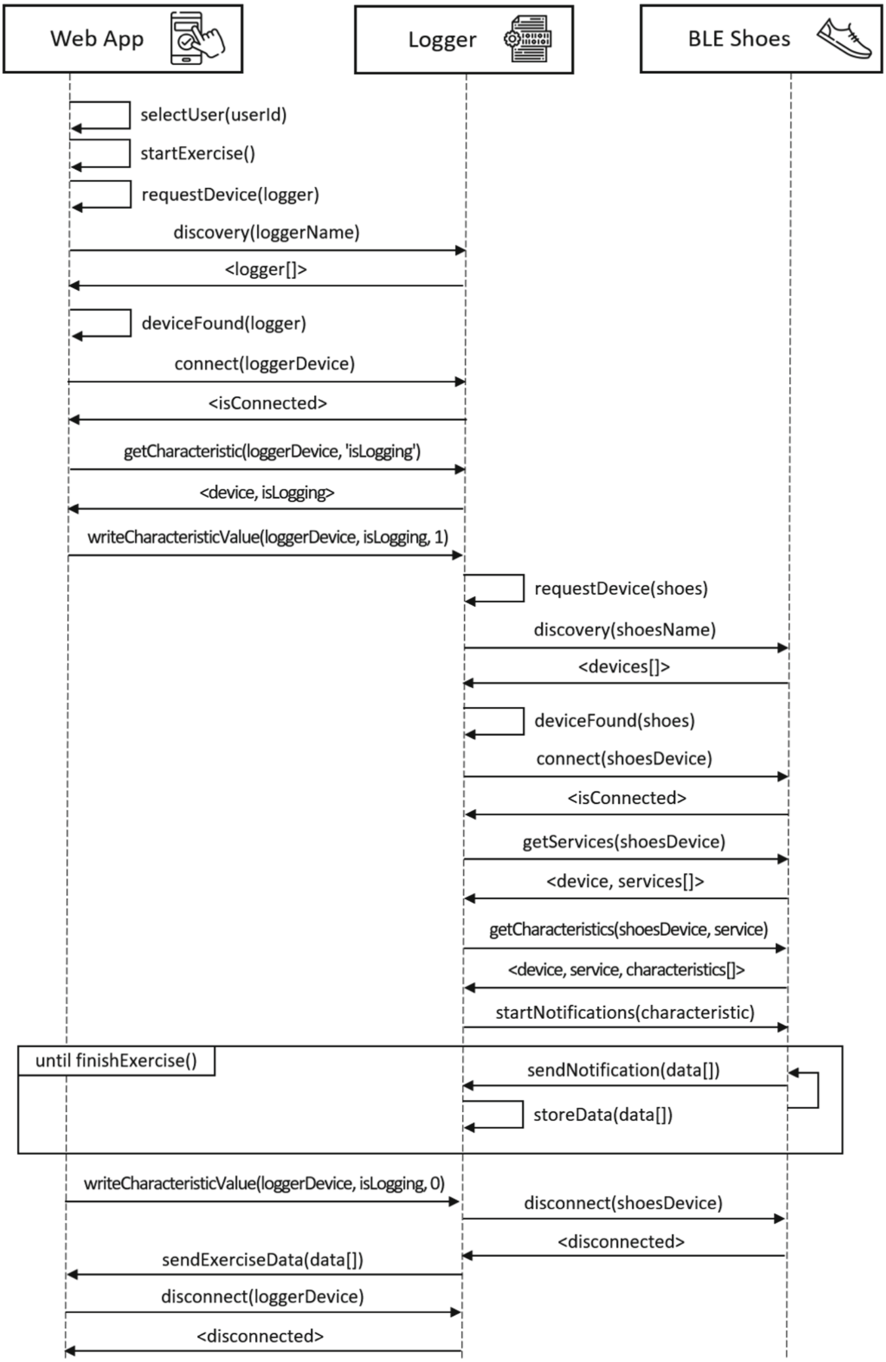


Fig. 6. Sequence diagram of the interaction among the devices in the scenario 2

The main components interactions taking place in this scenario are shown in Fig. 6. Once the clinician selects the user and starts the exercise, the Web App requests the discovery of the Bluetooth devices looking for a logger with a predefined name. If the logger is found, the Web App connects to it and searches for a specific characteristic used as a Boolean value to communicate whenever the logger must start or to stop the recording from the shoes. Writing the value ‘1’ to this characteristic, triggers the logger into discovering the shoe devices, connecting to them, searching for the services and characteristics, and starting the notification to receive the sensors data. During the exercise, the acquired data is locally stored on the logger device. As soon as the clinician stops the exercise from the web interface, the Web App writes the value ‘0’ to the logger characteristic causing its disconnection from the shoes and the transmission of the cached exercise data to the Web App. Eventually, when the data transfer in completed, the Web App disconnects from the logger device.

This configuration allows us to obtain the maximum performance in terms of connection reliability and transfer rate even in environments where several Bluetooth devices are present. Abstracting from the specific use case, this architecture is valuable whenever due to technical limitations or particular Bluetooth requirements, the connection between the Bluetooth equipment and the mobile device doesn’t perform well and can indeed be improved by employing external or multiple antennas. On the downside, with respect to the scenario 1, the on-site installation of an additional device increases the deployment complexity and might require protracted technical support during the system operation. Adopting a local Bluetooth connection between the mobile device on which the Web App is running and the logger device, avoids the need to set up a communication channel based on the Wi-Fi connection. Since in clinical environments it might be difficult to obtain access to existing Wi-Fi networks for security reasons, this would require to either deploy a local access point or a hotspot. Instead, using the Bluetooth connectivity, the user experience of the healthcare staff operating the Web App is not burdened with troublesome configuration operations.

4 Conclusions

In this paper, we presented an IoT solution to help in the clinical diagnosis of subjects with NMD by using gait information from a BLE sensorized shoe. Although we focused on collecting the data from the sensorized shoes to analyze the human gait characteristics, the IoT technological solutions we envisaged can be applied to any kind of BLE device that exposes an accessible data interface. This is possible thanks to the capabilities offered by the novel Web Bluetooth API that is a candidate to become the standard de facto for connectivity between smart devices and web applications. The presented architectural choices can be easily modified and adopted by any developer in need for a seamless integrated solution for their reference domain.

References

1. Engel, W.K.: Classification of neuromuscular disorders. Birth Defects Orig. Artic. Ser. **7**(2), 18–37 (1971). PMID: 4950913

2. Olsen, D.B., Gideon, P., Jeppesen, T.D., et al.: Leg muscle involvement in facioscapulohumeral muscular dystrophy assessed by MRI. J. Neurol. **253**, 1437–1441 (2006)
3. Gijsbertse, K., Goselink, R., Lassche, S., et al.: Ultrasound imaging of muscle contraction of the tibialis anterior in patients with facioscapulohumeral dystrophy. Ultras. Med. Biol. **43**(11), 2537–2545 (2017)
4. Dorobek, M., Szmidt-Sałkowska, E., Rowińska-Marcińska, K., Gaweł, M., Hausmanowa-Petrusewicz, I.: Relationships between clinical data and quantitative EMG findings in facioscapulohumeral muscular dystrophy. Neurol. Neurochir. Pol. **47**(1), 8–17 (2013)
5. Veltsista, D., Chroni, E.: Ultrasound pattern of anterolateral leg muscles in facioscapulohumeral muscular dystrophy. Acta Neurol. Scand. **144**(2), 216–220 (2021)
6. Barsocchi, P., et al.: Detecting user's behavior shift with sensorized shoes and stigmergic perceptrons. In: 2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT), pp. 265–268 (2019)
7. Barsocchi, P., Bianchini, M., Crivello, A., La Rosa, D., Palumbo, F., Scarselli, F.: An unobtrusive sleep monitoring system for the human sleep behaviour understanding. IEEE International Conference on Cognitive Infocommunications (CogInfoCom), pp. 91–96 (2016)
8. Varshney, G., Misra, M.: Push notification based login using BLE devices. In: 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), pp. 479–484 (2017)
9. Wåhslén, J., Lindh, T.: A javascript web framework for rapid development of applications in IoT systems for eHealth. In: 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 1–6 (2018)
10. Bardoutsos, A., Markantonatos, D., Nikoletseas, S., Spirakis, P.G., Tzamalis, P.: A human-centered Web-based tool for the effective real-time motion data collection and annotation from BLE IoT devices. In: 2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 380–389 (2021)
11. Stegman, P., Crawford, C.S., Andujar, M., Nijholt, A., Gilbert, J.E.: Brain–computer interface software: a review and discussion. IEEE Trans. Hum.-Mach. Syst. **50**(2), 101–115 (2020)

Compliance and Usability of an Asthma Home Monitoring System

Kevin C. H. Tsang[1](✉), Hilary Pinnock[1], Andrew M. Wilson[2], Dario Salvi[3], Carl Magnus Olsson[3], and Syed Ahmar Shah[1]

[1] Usher Institute, The University of Edinburgh, Edinburgh EH16 4UX, UK
`k.c.h.tsang@sms.ed.ac.uk`
[2] Norwich Medical School, The University of East Anglia, Norwich NR4 7TJ, UK
[3] Internet of Things and People, Malmö University, 211 19 Malmö, Sweden

Abstract. Asthma monitoring is an important aspect of patient self-management. However, due to its repetitive nature, patients can find long-term monitoring tedious. Mobile health can provide an avenue to monitor asthma without needing high levels of active engagement, and instead rely on passive monitoring. In our recent AAMOS-00 study, we collected mobile health data over six months from 22 asthma patients using passive and active monitoring technology, including smartwatch, peak flow measurements, and daily asthma diaries.

Compliance to smartwatch monitoring was found to lie between the compliance to complete daily asthma diaries and measuring daily peak flow. However, some study participants faced technical issues with the devices which could have affected the relative compliance of the monitoring tasks.

Moreover, as evidenced by standard usability questionnaires, we found that the AAMOS-00 study's data collection system was similar in quality to other studies and published apps.

Keywords: Asthma · Mobile Health · mHealth · Home Monitoring · Compliance · Passive Monitoring

1 Introduction

Asthma is a variable long-term condition affecting 339 million people worldwide [1]. There are often diurnal, seasonal, and life-time variations in the symptoms experienced by a patient. Common symptoms include shortness of breath, wheezing, and cough. Asthma attacks, if not treated promptly, can lead to hospitalization or even death [2, 3]. Currently, there is no cure for asthma, so the focus is on patients' self-management of their condition [4]. This involves monitoring asthma status to inform the best course of action.

Regular monitoring of asthma symptoms may identify worsening asthma status early and action can be taken to avoid further deterioration. However, patients may consider long-term monitoring as tedious, especially during extended periods when they do not experience symptoms, which may lead to a loss of engagement [5].

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 116–126, 2023.
https://doi.org/10.1007/978-3-031-34586-9_9

Mobile health (mHealth) can support asthma home monitoring and asthma self-management through the use of devices such as smartwatches which require much lower levels of active engagement from patients [6]. This ‘passive’ approach has the potential to support many more patients in monitoring and making decisions about their health.

We are currently working towards building a system to monitor asthma without requiring high levels of active engagement, the Asthma Attack Management Online System (AAMOS). It is currently unclear whether passive monitoring would be beneficial or indeed provide higher levels of engagement and compliance in long-term monitoring. As a starting point, we conducted the AAMOS-00 study “Predicting asthma attacks using connected mobile devices and machine learning”, a pilot study focused on collecting novel monitoring data (see [7] for additional details on the study design). The novel data collected during the AAMOS-00 study provides an opportunity to explore compliance with passive monitoring, the focus of this paper.

The primary aim of this study is to test whether passive monitoring would lead to higher compliance over an extended time when compared to active monitoring. The secondary aim is to investigate the usability of the system.

2 Methods

2.1 Study Design

We recruited 32 asthma patients across the United Kingdom who had experienced at least one severe asthma attack (as defined by the American Thoracic Society and European Respiratory Society [2]) during the past year. We undertook the observational study from April 2021 to June 2022 in two phases. During phase one, monitoring was by daily questionnaire over one month to select patients with at least 50% compliance. These patients (n = 22) were then invited to participate in phase two that consisted of device and daily questionnaire monitoring over six months. In addition to using their own smartphone, participants were provided with three smart monitoring devices: a smartwatch (MiBand3 by Xiaomi [8]), a smart peak flow meter (Smart Peak Flow Meter by Smart Asthma [9]), and a smart inhaler (FindAir ONE by FindAir [10]) [7]. Figure 1 provides an overview of the whole research data collection system.

In phase two, participants answered questionnaires at home using the Mobistudy app (a mobile-based research platform for data collection [11]). Participants uploaded data from the smartwatch weekly via a Bluetooth connection, and conducted two sets of peak flow measurements with the smart peak flow meter, once in the morning and once at night; the maximum of three measurements was reported per set. Moreover, participants used the FindAir app to upload data from the smart inhaler. At the end of the AAMOS-00 study, participants were asked to digitally complete an exit questionnaire at home about the acceptability and usability of the study’s data collection system.

Our analysis used data collected from the 22 participants during phase two of the AAMOS-00 study and the end-of-study questionnaire. In particular, the focus was on investigating passive and active monitoring using the daily asthma diary, smart peak flow meter, and smartwatch usage data. The daily asthma diary and smart peak flow meter monitoring tasks reflected current practice of asthma monitoring, while smartwatch use represented a promising technology for passive monitoring.

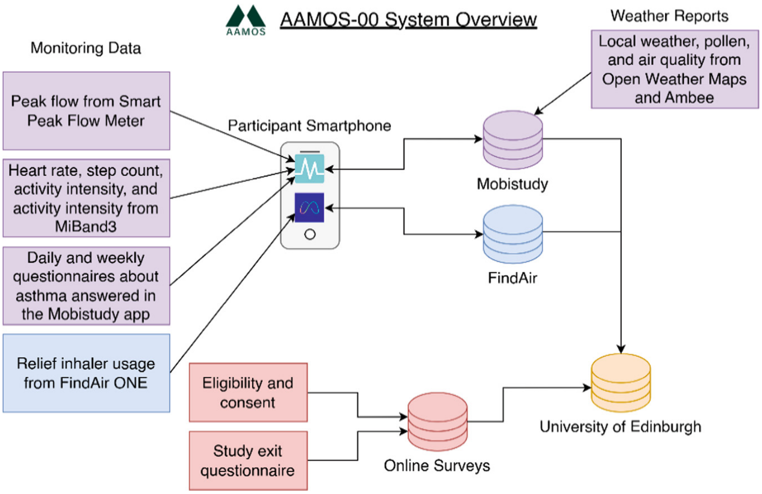

AAMOS
AAMOS-00 System Overview
Monitoring Data
Peak flow from Smart Peak Flow Meter
Heart rate, step count, activity intensity, and activity intensity from MiBand3
Daily and weekly questionnaires about asthma answered in the Mobistudy app
Relief inhaler usage from FindAir ONE
Participant Smartphone
Weather Reports
Local weather, pollen, and air quality from Open Weather Maps and Ambee
Mobistudy
FindAir
Eligibility and consent
Study exit questionnaire
Online Surveys
University of Edinburgh

Fig. 1. AAMOS-00 system overview (adapted from our previous publication [7]).

2.2 Measure of Compliance

Compliance with each monitoring task was defined as their completion on a daily basis. Specifically, daily compliance with active monitoring meant completion of the asthma diary (a seven-item questionnaire) daily task and completion of at least one set of peak flow readings (three readings per set) per day. Daily compliance for the passive monitoring task meant wearing the smartwatch for at least 12 h between 00:00 and 23:59. The mobile app provided daily notification reminders to complete the monitoring tasks.

The day 0 compliance was the daily compliance on the first day of data collection. The average compliance over each month was calculated across the study population, which was the total tasks completed over 30 days by all participants divided by the total engagement requested. For example, the average compliance of the asthma diary in the first 30 days for 22 participants was

$$\frac{\text{Total asthma diaries completed in 30 days by 22 participants}}{30 \times 22} \tag{1}$$

Change in compliance over time was investigated using linear regression using R, which gave an intercept and gradient per monitoring task. This intercept represented the initial level of compliance, and the gradient represented the average increase or decrease in compliance over 30 days.

Participant retention was defined to be the total number of days between the first and last day of engagement with the study. Participants in phase two of AAMOS-00 each had a potential maximum of 184 days of participation.

2.3 Usability Questionnaires

The AAMOS-00 study exit questionnaire about the acceptability and usability of the system incorporated three validated questionnaires. Usability was assessed with the

System Usability Scale (SUS) [12], personal motivation to use technology for self-management used the mHealth Technology Engagement Index (mTEI) [13], and app quality and perceived impact used the User version of Mobile Application Rating Scale (uMARS) [14]. Some uMARS questions were adapted to reflect the AAMOS-00 study's aims and system.

System Usability Scale (SUS). The SUS [12] questionnaire is a widely-used 10-item validated questionnaire [15] assessing system usability. It is answered on a five or seven point Likert scale. The questions are simple and effective [15], and alternate between positively and negatively worded text, to reduce response bias [12].

mHealth Technology Engagement Index (mTEI). The mTEI [13] is a 16-item validated questionnaire which measures a person's motivation to use telehealth systems by asking about five main areas: autonomy, competence, relatedness, goal attainment, and goal setting. Assessing the correlation to other related measures, the mTEI developers found significant positive correlations [13] with the Psychosocial Impact of Assistive Devices (PIADS) [16], but low correlations to the Technology Acceptance Model (TAM) [17], and SUS [12], suggesting they were distinct measures [13]. We considered all the questions to be helpful in assessing a person's self-management status and preferences.

User Version of Mobile Application Rating Scale (uMARS). The uMARS [14] validated questionnaire builds upon the MARS [18] which measures app quality. The questionnaire includes 16 questions in four domains: engagement, functionality, aesthetics, and information, and two sections on subjective quality and perceived impact. The answers include five statements along a scale of "1. Inadequate" to "5. Excellent". To be consistent with the other two questionnaires (SUS and mTEI), the uMARS questions were reworded to a five-point Likert scale format in the AAMOS-00 exit questionnaire.

3 Results

3.1 Study Population

Most participants in phase two of the AAMOS-00 study were female (77%), white (95%), and had uncontrolled asthma in the month before joining the study (see Table 1). The average age was 40 years, and all the participants in phase two of the AAMOS-00 study had at least 50% compliance in phase one (one month of daily questionnaire completion).

3.2 Compliance

The compliance to monitoring did not show significant difference between the 'passive' smartwatch and 'active' monitoring tasks – all were equally low at <50% by the end of the second month. The highest compliance was to the asthma diary task, which started with 82% compliance on day 0 and continued to have the highest compliance throughout the six months. Compliance to peak flow monitoring was the lowest, beginning at 46% on day 0 and dropping to 16% after six months. The compliance to smartwatch monitoring

Table 1. Population Characteristics. Twenty-two patients participated in phase two of the AAMOS-00 study, where participants conducted six months of monitoring using smart devices and answered daily questionnaires about asthma. RCP3 score ranges from 0 to 3, 0 indicating good control, 3 indicating poor control [19].

Characteristics	AAMOS-00 Phase Two (n = 22)
Sex, n (%)	
Female	17 (77%)
Male	5 (23%)
Age, median (IQR)	40.2 years old (15.7 years old)
Royal College of Physicians' "3 Questions" about asthma control (RCP3) [19] **in past month**, mean	2.4

was in-between the compliance levels of asthma diary and peak flow monitoring in all six months (see Fig. 2).

Furthermore, by investigating the linear fit of the change in compliance over six months, we observed the level of compliance to smartwatch monitoring was between the two active monitoring tasks. Although the asthma diary started with the highest level of compliance, it also had the largest drop in compliance per month (−7.6% compliance per 30 days). In contrast, the peak flow task started with the lowest compliance but also had the lowest drop in compliance per month (−4.9% compliance per 30 days). See Table 2.

Table 2. Linear fit of compliance over 6 months.

Monitoring Task	Intercept	Average change in compliance per 30 days (gradient)
Asthma Diary	62%, 95% CI [55%, 69%], $t = 17$, $p = 6.7e{-}05$	−7.5%, 95% CI [−9.6%, −5.5%], $t = -7.2$, $p = 0.0019$
Smartwatch	51%, 95% CI [46%, 57%], $t = 18$, $p = 6.1e{-}05$	−6.3%, 95% CI [−7.9%, −4.6%], $t = -7.4$, $p = 0.0018$
Peak Flow	42%, 95% CI [38%, 46%], $t = 22$, $p = 2.5e{-}05$	−4.9%, 95% CI [−5.9%, −3.8%], $t = -8.9$, $p = 0.00088$

3.3 Questionnaire Feedback

More than half of the phase two participants (14 out of 22) filled in the end of study questionnaire. However, this small sample of respondents was skewed towards participants who were very adherent to monitoring, even when compared to the study population (in themselves motivated individuals). Respondents to the final questionnaire had averaged 154 days of participation which was higher than the overall average in phase two

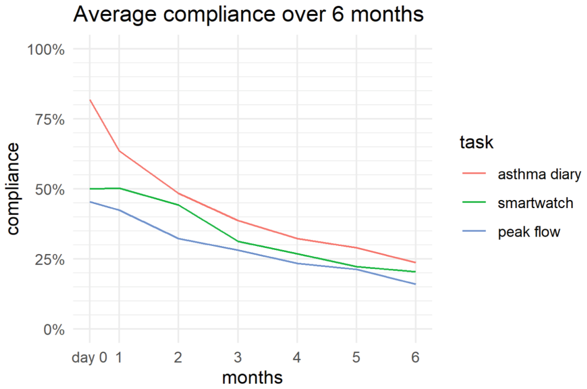


Fig. 2. Compliance to monitoring in the AAMOS-00 study. Twenty-two asthma patients were asked to complete daily monitoring tasks. Compliance was measured by completion of daily asthma diary, wearing the smartwatch at least 12 h per day, and conducting a set of peak flow measurements per day.

(123 days). The two respondents with the lowest retention had five and 38 days of participation. The respondent with five days of participation formally withdrew from the study citing frustration with the technology. Median age of respondents was 47 years old (slightly older than the overall phase two study population) with 71% females (slightly smaller proportion of females compared to the overall phase two study population).

3.4 Questionnaire Score

Median SUS score in the AAMOS-00 study was 61.25, which is slightly below the average SUS score of 68 as measured across 500 studies [20]. The median overall uMARS score was 3.44, which is slightly higher than the average score of 3.26 as measured across 50 mental health and well-being apps on the iTunes store [18].

Investigating the uMARS score further, we could see the lowest scored aspects were engagement (AAMOS-00 median score of 3.20, which is still higher than the iTunes average of 2.68 [18]) and functionality (AAMOS-00 median score of 3.5, which is lower than the iTunes average of 4.01 [18]) (see Fig. 3). The two highest scored aspects of aesthetics (median score of 3.67) and information (median score of 3.75) were higher than the iTunes average of 3.49 and 2.88 respectively [18].

The number of asthma diaries completed gives an approximate measure for the engagement with the study. There was a strong correlation (Pearson's correlation = 0.56) between SUS score and total asthma diaries completed, which suggests that the usability of the system was a major factor influencing engagement with the study.

In general, there was a weak correlation (Pearson's correlation = 0.28) between the mTEI score and the total asthma diaries completed, suggesting that motivation to use technology was mostly independent of engagement. However, the autonomy subfactor of the mTEI questionnaire had a moderate correlation (Pearson's correlation = 0.34) with engagement, indicating users who are motivated by a need to be in control of their own health were more likely to complete asthma diaries.

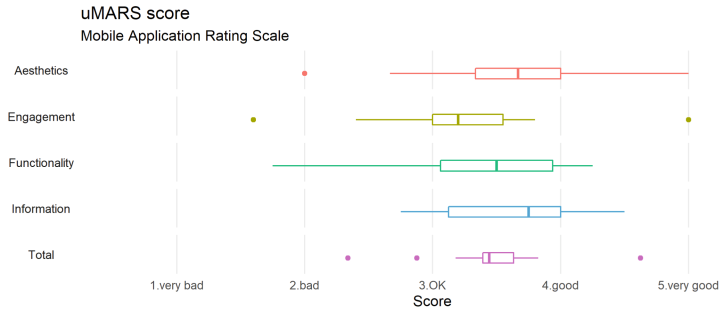


Fig. 3. uMARS score of the AAMOS-00 study's data collection system. The median uMARS overall score was 3.44.

There was moderate correlation (Pearson's correlation = 0.32) between the number of asthma diaries completed and the uMARS questionnaire score. In particular, there was a moderate correlation (Pearson's correlation = 0.43) with the functionality section, and a low correlation (Pearson's correlation = 0.24) with the information section of the uMARS questionnaire. This suggests that participants who found the research data collection system easy to use and considered that it provided useful and reliable information were more likely to engage with the study.

3.5 Study Feedback

The devices (smartwatch, smart peak flow meter, and smart inhaler) were generally reliable, but some participants encountered issues with different devices during the study. One participant pointed out that the readings of the smart peak flow meter did not match with their mechanical counterpart, a recognized discrepancy that could hinder clinical adoption of the device. Although the hardware designers of the smart inhaler device had tackled some problems with false positives and false negatives, there were still comments about the smart inhaler's reliability. Additionally, some people encountered missed actuations (false negatives). Some smartwatches also had to be replaced after around five months of use, when the device stopped holding charge or failed to connect to the app via Bluetooth.

> *"The FindAir app was the most useful tracking inhaler usage but needs to be more reliable. It kept missing uses."*
>
> *– participant (female, 38 years old, 183 days of participation)*
>
> *"The smart peak flow meter was not recording at the same reading as the regular more traditional widely used peak flow tube issued by GP's and the pharmacy."*
>
> *– participant (female, 47 years old, 184 days of participation)*
>
> *"The peak flow meter did not always work and it became frustrating to use"*
>
> *– participant (female, 37 years old, 184 days of participation)*
>
> *"The FindAir app never worked for me; the peak flow meter occasionally didn't work and the [smartwatch] had to be replaced."*

> *– participant (female, 46 years old, 184 days of participation)*
>
> *"I have a iPhone 11 [and Apple] watch 3. Both are far more advanced and could do a better job more accurately and reliably"*
>
> *– participant (male, 52 years old, 5 days of participation)*

Although the smart peak flow meter worked well in controlled environments, it had trouble calibrating with some LED lights when used in a real-world setting in this study. The flickering lights would sometimes drastically inflate the peak expiratory flow (PEF) rates to impossible values. A few participants who could not use the smart peak flow meter reliably used other peak flow measurement methods and manually shared their PEF recordings via email.

> *"[The peak flow meter] doesn't work in normal indoor light settings which made it harder to use in autumn and winter when day light hours are limited. … Gave drastically inaccurate readings occasionally."*
>
> *– participant (female, 47 years old, 184 days of participation)*
>
> *"Downside is that the peak flow doesn't work with LED lighting (lightbulbs we have in UK)"*
>
> *– participant (female, 48 years old, 184 days of participation)*

The AAMOS-00 study was an observational study and did not actively provide any medical advice, but some participants found the monitoring alone to be useful. This included, for example, seeing the disparity between the measured relief inhaler usage and their own answers to the question about daily relief inhaler usage.

> *"I was surprised by how out I was when guessing how many times I'd used my inhaler."*
>
> *– participant (female, 54 years old, 184 days of participation)*
>
> *"Thank you for sending me [the FindAir] device as it has opened my eyes up to how much stress affects my asthma and is a big trigger for me. It made me realize that I need to be more aware of this and take more action."*
>
> *– participant (female, 47 years old, 184 days of participation)*

In contrast, some respondents did not think the study and monitoring had changed their attitudes toward improving their asthma.

> *"I was very happy to record the data, but did not find it helped me to manage my asthma."*
>
> *– participant (female, 46 years old, 184 days of participation)*

During the study, some participants encountered multiple issues with the technology (e.g. setting up the Bluetooth connection between Mobistudy and the smartwatch). We resolved most software problems via emails and video calls with participants, but some issues were escalated to the Mobistudy (provider of the main system for data collection), Smart Asthma (provider of the smart peak flow meter), and FindAir (provider of the smart inhaler) technical team. Hardware issues were resolved sometimes by detailed instructions or by sending replacements.

> *"I would like thank Kevin Tsang for his rapid and patient help when devices didn't work."*
>
> *– participant (female, 46 years old, 184 days of participation)*
>
> *"Thank you for the support when issues did arise."*
>
> *– participant (female, 37 years old, 184 days of participation)*

4 Discussion

We have found no evidence that a passive monitoring task (wearing the smartwatch) provided a higher level of engagement when compared to active monitoring tasks (completing asthma diaries and taking peak flow measurements) used in current practice of asthma self-management. The compliance to monitoring with the smartwatch was between compliance level to monitoring with the asthma diary and the smart peak flow meter – and both fell off rapidly so that by the end of six months only a quarter of people were still monitoring. This result could be confounded by the technical issues with the devices, because the asthma diary task had minimal technical issues, whereas several participants encountered issues when using the smart monitoring devices.

Feedback from users revealed the challenges with the three monitoring devices, especially with taking peak flow measurements. Although the mobile asthma diary task had fewer technical issues, there was a relatively similar levels of compliance (10%–20% difference) between the daily diary and the daily smart peak flow meter measurements suggesting that the participants were highly motivated to handle technical issues. Overall, the technology would need to be more reliable before it could be widely adopted.

Furthermore, due to the technical implementation of the smartwatch data collection, it required some active engagement from users to upload the smartwatch data weekly to their smartphone which may have been a significant disincentive. This limited our exploration of the potential for fully 'passive' monitoring requiring no effort on the part of the user once it has been set up. Technical issues and smartwatch implementation may have led to lower compliance than expected [21, 22]. Moreover, some patients already owned and regularly used a smartwatch, which may have affected their willingness to use a secondary (likely less sophisticated) device for the study.

When compared to other studies and published apps, the AAMOS-00 study's data collection system was similar in quality, evidenced by standard questionnaires SUS and uMARS. However, the small number of respondents were likely to have been skewed toward highly motivated participants who found the system more usable as they had a higher average retention compared to the overall study population. The usability scores should be interpreted considering this possible bias.

Another limitation was the narrow selection criteria, which selected asthma patients who had an interest in monitoring and had experienced a severe asthma attack in the past 12 months, yielding a small sample size of the AAMOS-00 study. The average retention in phase two (which included daily tasks) was 123 days. This is similar to the average of 122 days patients have been willing to engage in previous studies [7, 22–24]. However, it is plausible to expect a substantially lower level of retention in the wider population. During a patient and public involvement focus group session that we undertook, some patients suggested that the retention amongst the wider population could be as low as one week.

There are still areas of unexplored questions within the AAMOS-00 dataset. Our future work includes deeper analysis to investigate each device through correlating themes in user feedback with compliance data. Future studies could consider investigating the effect of reminders and other interventions (e.g. improved feedback and gamification strategies) to increase compliance over an extended time and explore the nuanced barriers of each monitoring task. Additionally, future studies may consider extracting data from the devices patients may already be using.

5 Conclusions

In the AAMOS-00 study, a small-scale study conducted with highly motivated patients, the compliance to passive (smartwatch) and active (daily asthma diary and peak flow measurement that are currently used in asthma self-management) monitoring was similar. Although the AAMOS-00 study faced some technical issues, the quality of the data collection system was comparable to other studies and published apps and is a promising option for future mHealth studies.

Acknowledgements. This work was supported by Asthma + Lung UK as part of the Asthma UK Centre for Applied Research grant number AUK-AC-2018-01. We thank all the participants of the AAMOS-00 study, this research would not be possible without their time and effort. We thank the AUKCAR PPI for their involvement in analyzing the data. We thank the Mobistudy team and Malmö University for their support with data collection. Their work was supported by the Knowledge Foundation, through the Internet of Things and People research profile.

References

1. Global Asthma Network: The Global Asthma Report 2018. Global Asthma Network, Auckland (2018)
2. Reddel, H.K., Taylor, D.R., Bateman, E.D., et al.: An official American thoracic society/European respiratory society statement: asthma control and exacerbations - standardizing endpoints for clinical asthma trials and clinical practice. Am. J. Respir. Crit. Care Med. **180**, 59–99 (2009)
3. Global Initiative for Asthma (GINA): Global Strategy for Asthma Management and Prevention (2021)
4. Pinnock, H.: Supported self-management for asthma. Breathe **11**, 98–109 (2015)
5. May, C.R., Montori, V.M., Mair, F.S.: We need minimally disruptive medicine. BMJ **339**, 485–487 (2009)
6. Tsang, K.C.H., Pinnock, H., Wilson, A.M., Shah, S.A.: Application of machine learning algorithms for asthma management with mHealth: a clinical review. J. Asthma Allergy **15**, 855–873 (2022)
7. Tsang, K.C.H., Pinnock, H., Wilson, A.M., et al.: Predicting asthma attacks using connected mobile devices and machine learning: the AAMOS-00 observational study protocol. BMJ Open **12**, e064166 (2022)

8. Xiaomi Xiaomi UK. https://www.mi.com/uk/. Accessed 9 Mar 2022
9. Smart Asthma Smart Asthma. https://smartasthma.com/. Accessed 9 Mar 2022
10. FindAir FindAir. https://findair.eu/. Accessed 9 Mar 2022
11. Salvi, D., Magnus Olsson, C., Ymeri, G., et al.: Mobistudy: mobile-based, platform-independent, multi-dimensional data collection for clinical studies. In: 11th International Conference on the Internet of Things, pp. 219–222. ACM, St.Gallen (2021)
12. Brooke, J.: SUS: a "quick and dirty" usability scale. In: Usability Evaluation in Industry, pp 207–212. CRC Press (1996)
13. Dewar, A.R., Bull, T.P., Malvey, D.M., Szalma, J.L.: Developing a measure of engagement with telehealth systems: the mHealth technology engagement index. J Telemed. Telecare **23**, 248–255 (2017)
14. Stoyanov, S.R., Hides, L., Kavanagh, D.J., Wilson, H.: Development and validation of the user version of the mobile application rating scale (uMARS). JMIR mHealth uHealth **4**, e72 (2016)
15. Brooke, J.: SUS: a retrospective. J. Usabil. Stud. **8**, 29–40 (2013)
16. Demers, L., Monette, M., Descent, M., et al.: The psychosocial impact of assistive devices scale (PIADS): translation and preliminary psychometric evaluation of a Canadian-French version. Qual. Life Res. **11**, 583–592 (2002)
17. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. Manag. Inf. Syst. **13**, 319–339 (1989)
18. Stoyanov, S.R., Hides, L., Kavanagh, D.J., et al.: Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth **3**, e27 (2015)
19. Pearson, M., Bucknall, C.: Measuring Clinical Outcome in Asthma: A Patient-Focused Approach. R Coll Physicians, London (2000)
20. Sauro, J.: A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices. CreateSpace Independent Publishing Platform, Denver (2011)
21. Xu, X., Tupy, S., Robertson, S., et al.: Successful adherence and retention to daily monitoring of physical activity: lessons learned. PLoS One **13**, e0199838 (2018)
22. Compernolle, S., Vandelanotte, C., Cardon, G., et al.: Effectiveness of a web-based, computer-tailored, pedometer-based physical activity intervention for adults: a cluster randomized controlled trial. J. Med. Internet Res. **17**, e38 (2015)
23. Hui, C.Y., McKinstry, B., Walton, R., Pinnock, H.: A mixed method observational study of strategies to promote adoption and usage of an application to support asthma self-management. J. Innov. Heal Inform. **25**, 243–253 (2019)
24. Senturia, Y.D., McNiff Mortimer, K., Baker, D., et al.: Successful techniques for retention of study participants in an inner-city population. Control Clin. Trials **19**, 544–554 (1998)

Assessing Older Adult's Gait Speed with Wearable Accelerometers in Community Settings: Validity and Reliability Study

Antonio Cobo[1,2](✉), Elena Villalba-Mora[1,2], Rodrigo Pérez-Rodríguez[3],
Juan E. Medina[1], Paula Robles-Mateos[1], Ángel Rodríguez-Laso[4],
and Leocadio Rodríguez-Mañas[4,5]

[1] Centre for Biomedical Technology (CTB),
Universidad Politécnica de Madrid (UPM), Madrid, Spain
{antonio.cobo,elena.villalba}@ctb.upm.es
[2] CIBER de Bioingeniería Biomateriales y Nanomedicina,
Instituto de Salud Carlos III, Madrid, Spain
[3] Fundación para la Investigación Biomédica del Hospital Universitario de Getafe,
Hospital de Getafe, Madrid, Spain
rodrigo.perez@urjc.es
[4] CIBER de Fragilidad y Envejecimiento Saludable, Instituto de Salud Carlos III,
Madrid, Spain
{arodriguezlaso,leocadio.rodriguez}@salud.madrid.org
[5] Servicio de Geriatría, Hospital de Getafe, Madrid, Spain

Abstract. We present the preliminary results of a validity and reliability study of two different state-of-the-art algorithms to estimate the gait speed of older adults in free-living conditions from the data collected by the ActiveUP wearable device. We described the ActiveUP wearable sensor together with its integration in a smart environment for frail and pre-frail older adults via an edge-computing architecture. A cross-sectional observational study was conducted. A sample of 18 people, 77.89 (6.47) y.o. 11 women, was recruited and their movement signals were recorded during short (2.4 m) and long (6 m) walking bouts. Validity, agreement, and reliability were assessed with Pearson's correlation coefficient, with SEM and Bland-Altman limits of agreement (LOA), and with an ICC(A, 1) model of the intra-class correlation coefficient, respectively. Validity and reliability seemed to be good. However, the small size of our sample results in broad confidence intervals for the estimators. The agreement seems not to be good enough to trigger therapeutic responses. More studies are necessary to test whether the threshold for clinical tests is applicable under free-living conditions.

Keywords: gait speed · wearabe accelerometer · validity · reliability

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 127–137, 2023.
https://doi.org/10.1007/978-3-031-34586-9_10

1 Introduction

The World Health Organization defines healthy aging as “the process of developing and maintaining functional ability that enables well-being in older age” [1] (p. 28). This definition emphasizes that healthy aging is more than the simple absence of disease. It is now widely recognized that functional impairment, rather than disease, is the main risk factor for disability and death [2] and that it is the main factor that explains the increase in cost for social and health systems. Older adults with functional impairment and their families make an intensive use of health and social services [3]; and this presents a major challenge in providing the health services that older people need while maintaining the sustainability of the system. The path to disability is a gradual process of functional loss [4]. During this process, even years before developing a disability, older adults show characteristic signs of a syndrome known as frailty [4]. Frailty is a state of increased vulnerability to adverse outcomes due to a reduction in the ability to respond to stressors, even if they are of low intensity [5]. However, frailty can be prevented and also reversed [6]. In fact, there are validated tools to detect frailty and effective interventions to manage it [5]. Frailty is evaluated by trained geriatricians or geriatric nurses in specialized care. However, specialized care does not have enough resources to screen the entire older population; the identification of new and more efficient forms of detection and screening remains a challenge [7].

The use of wearable automatic sensors has been proposed as a way to assess the functional status of older people without involving specifically trained personnel [8]. Gait analysis is one of the most studied applications of wearable sensors; and gait speed is one of the five markers of frailty in the most widely used frailty model, the Linda Fried phenotypic model [4]. Gait sensors have been used to implement instrumented versions of standard clinical tests. For example, they have been used to estimate the value of gait speed and some other kinematic variables in walking tests of different lengths [9,10]. The usability of wearable sensors for instrumented walking tests has not been tested in older adults in unsupervised home settings. However, the favorable experience of Cobo et al. with their body sensor for instrumented sit-to-stand tests [11] suggests that instrumented walking tests with wearables could be suitable for this scenario.

Gait sensors have also been used to go a step further and estimate gait speed without interfering with people’s daily activities. Recent studies using wearable accelerometers to collect gait signals under free-living conditions can be found in the scientific literature [12–14]. The wearable sensors described in these studies were located on the waists of the participants or on their lower backs through elastic or adjustable belts. Their algorithms processed the acceleration signals to identify sustained walking bouts and then provided an estimate of gait speed for each bout. They report good results. However, there are still some challenges to overcome. On the one hand, although a relationship between daily gait speed measurements and functional impairment has already been observed [10,15], more studies are still needed to find out which estimation methods best capture the onset of functional changes and which quantitative thresholds best quantify their intensity. On the other hand, sensor measurements must be readily available to older adults' physicians to assess their functional status and make therapeutic decisions when appropriate. Smart environments can connect these gait speed sensors to third-party services via the Internet of Things (IoT) to provide personalized, anticipatory and adaptive services in many areas, such as energy management, health care, quality of life (independent and assisted living), or social isolation [16,17]. However, sensors in a smart environment require connectivity, which adds a load to the computation and energy constraints of the devices.

In the present paper, we describe the ActiveUP wearable sensor as an IoT device integrated in an edge-computing architecture; and present the preliminary results of a validity and reliability study of three different state-of-the-art algorithms to estimate gait speed in free-living conditions from the data collected by the ActiveUP wearable device.

2 Methodology

We conducted a prospective cross-sectional observational study to assess the validity and reliability of three different state-of-the-art algorithms for the estimation of gait speed under free-living conditions. A sample of older adults was recruited and their movement signals were recorded during short and long walking bouts 2.4 m and 6 m long, respectively. The study was carried out according to the Declaration of Helsinki and the protocol was approved by the Ethics Committee of the University Hospital of Getafe (CEIm21/41).

2.1 Participants

Subjects were eligible for the present study if they:

- met ALL the following INCLUSION CRITERIA:
  - subjects 70 years or older,
  - subjects able to walk, with or without mobility aids (such as canes or walkers).
- did not meet ANY of the following EXCLUSION CRITERIA:
  - subjects unwilling or unable to give their consent,
  - subjects unable to understand the researchers' commands or the questionnaires. This criterion was tested asking subjects to point to the device on/off button and to describe the meaning of the light that turned on after pressing it.
  - Clinically unstable subjects in the judgment of the investigator.

2.2 Apparatus

The ActiveUP Wearable Sensor. This sensor is a device of 10 × 7 cm that includes a 6 degree-of-freedom inertial measurement unit (IMU). Figure 1 shows three pictures of the sensor. The device comprises an ESP32 microcontroller with WiFi and BT interfaces onboard, an MPU-6050 GY-521 IMU with an I2C interface, a DS3231 Real Time Clock, a Micro-SD card module with an SPI interface, a TP4056 Lipo charger with a USB type-C interface, an RGB LED, and a commutator. The microcontroller samples linear acceleration and angular velocity 18 Hz in each of the three spatial directions and stores the data on an SD card for subsequent transmission.

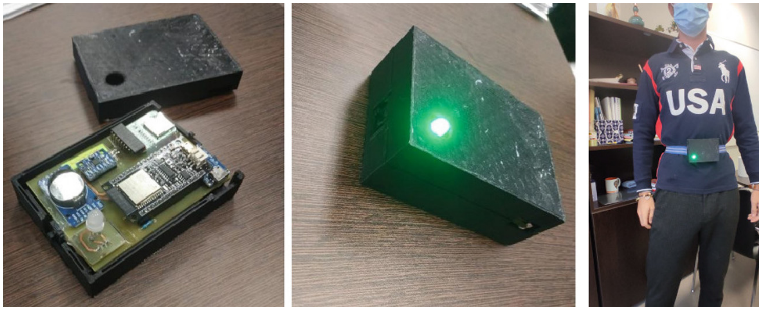


Fig. 1. The ActiveUP wearable sensor. View of the prototype PCB (left). View of the operative device (center). Subject wearing the device (right).

The ActiveUP wearable sensor has been integrated into an edge computing architecture as shown in Fig. 2. The architecture is a wireless network that comprises sensors, an edge node, and a gateway with an Internet connection. The edge node has been implemented on a Raspberry Pi and includes a message broker, local storage, and some data processing modules. The message broker redirects sensor data to authorized interested parties (in-home and external) via a publish/subscribe mechanism. Data processing modules transform raw data into clinically relevant information, thus reducing the amount of information that must be transmitted over the Internet.

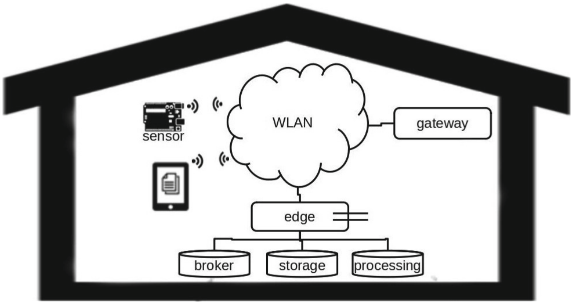


Fig. 2. Edge-computing architecture for the ActiveUP home system.

The Gait Speed Algorithms. We processed the signals collected with the ActiveUP sensor with three state-of-the-art algorithms to estimate the gait speed of older adults.

Mueller's Algorithm. Mueller et al. described an algorithm specifically designed for older adults who walk at a slow speed [13]. The algorithm uses a Short-Time Fourier Transform (STFT) to divide the acceleration signal on each axis into 2.5-s-long overlapped windows. Then, windows are kept or discarded based on the plausibility of their dominant frequency and trunk inclination. Finally, the algorithm applies a Hilbert transform on each axis and estimates gait speed by entering the amplitude values into a previously fitted linear regression model. They do not provide a precise description of some of the steps of the algorithm. Thus, we implemented our own adaptation by applying the following criteria:

1. we computed the STFT step by splitting the signal into 2.5-s windows with a 50% overlap (empirical value).
2. We used the highest frequency in a window as the dominant frequency, regardless of the axis.
3. We removed windows with a dominant frequency below 0.5 Hz (empirical value).
4. We did not use the angle of the trunk to remove any windows because it resulted in a decreased performance.
5. We did not apply the Butterworth filter in the sixth step because the filter parameters were not described.

GaitPy. GaitPy uses a pre-trained binary classifier to detect bouts. Then, it enhances the patterns in the vertical axis with a wavelet-based method to detect heel strikes and toe off events. Finally, the acceleration in the vertical axis is integrated to derive a vertical displacement and an inverted pendulum model is applied to estimate gait speed on a stride-to-stride basis [12]. It is publicly available as an open source Python package [18] at https://github.com/matt002/GaitPy. GaitPy requires sampling frequencies 50 Hz for the input signals. Since the sampling frequency in our device 18 Hz, we had to up-sample the signals with an interpolation filter before applying GaitPy. We used the interp function in Matlab R2022a with a multiplication factor of 3.

Urbanek's Algorithm. Urbanek et al. divide the acceleration signal of each axis with an STFT and rely on the harmonic nature of sustained walking to quantify the local periodicity of the signal within each window. Then, they estimate the fundamental frequency of the observed signals and identify bouts of sustained walking by grouping together windows with low variability of step frequency. Finally, they provide an estimate of cadence (that is, step frequency) for each bout [10]. They thoroughly described their algorithm; therefore, we were able to code our own implementation in Matlab. However, the algorithm has a couple of drawbacks. They report optimal results for a minimum bout duration of 10 s; therefore, performance is expected to degrade for short bouts (2.4 m) and long bouts (6 m) at home. In addition, their algorithm provides estimations of cadence rather than estimations of gait speed. Thus, we only tested Urbanek's algorithm's ability to identify bouts (which is out of the scope of this paper), but did not test its validity and reliability as a gait-speed estimator.

Reference Measurements. The reference measurement of gait speed in short bouts was measured with our previously validated device for 2.4 m walking tests (2.4 mWT) [19]. This device comprises a foldable tape equipped with ultrasound sensors (Fig. 3) and measures the time it takes for a subject to walk along the tape (from the ultrasound sensor at the beginning to the ultrasound sensor at the end). Finally, it reports the average speed of the subject.

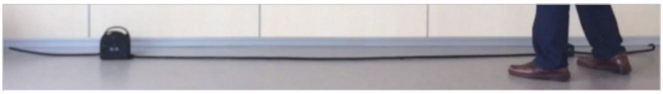


Fig. 3. Gait speed sensor for the 2.4 mWT.

In the case of long bouts, we obtained the reference measurements from a standard 6m walking test (6mWT) with a manual stopwatch.

2.3 Procedure

Participants were recruited from the Day Hospital and outpatient clinics of the Getafe University Hospital geriatric service and among relatives and acquaintances of the members of the research team. The day hospital health care personnel and those of external consultations asked patients, after their usual appointment, if they wanted to be informed about a study. Those who accepted went to a separate consultation in which a member of the research team individually explained the study and answered all the questions and doubts they had. In the case of relatives and acquaintances, the different members of the research team contacted them and asked them if they wanted to be informed about a study. They were individually contacted in person or by phone and received an explanation of the study and answers to all their questions and doubts. All of them, regardless of the entry mechanism, received enough time to assess their participation. The possibility of postponing the decision for several days was considered, which is why a contact phone number was provided to clarify doubts or arrange a later appointment for the study. The information related to the study was provided by the physician, nurse, or any other qualified member of the research team. When appropriate, the caregivers of the patients or their family members also received the information.

Subjects willing to participate were screened according to the eligibility criteria. The selected subjects signed an informed consent and were included in the study. Participants included in the study were asked for their consent to video record the data collection session. Those who accepted received an additional authorization sheet to sign. Those who did not accept continued to participate in the study without being recorded. Then the training session and the experimental data collection session took place. If the participant authorized the recording of the session, the recorded timespan was limited to the moments when the participant used the wearable sensor and only captured the image of the participant from the waist down.

During the training session, the participants learned how to operate the device to perform its function.

Data collection involved:

- a questionnaire for demographic and anthropometric data,
- wearing the sensor while taking two gait speed tests (2.4 mWT and 6 mWT),
- a short simulation of daily activities, out of the scope of the present paper,
- an interview about the participants' experience in the use of the technology and their impressions about the sensor,
- and, finally, wearing the sensor while taking two gait speed tests (2.4 mWT and 6 mWT) once again.

During the experiment, a member of the research team took note of the output of the reference measurements. After the data collection session, the collected signals were processed with the algorithms described above and the values of their output were added to the data set.

2.4 Analysis

To characterize the reliability of the measurements, the test-retest reliability was studied. Since gait speed is a continuous variable, the test-retest reliability was estimated by calculating intra-class correlation coefficients (ICC). An ICC (A,1) model was calculated with the icc function in the irr package in the statistical software R, version 4.2.1 [20].

The degree to which two measurements are identical (agreement) was estimated by calculating the standard error of the measurement (SEM). We used SEM estimations to estimate the minimum detectable difference (MDD) and compared it to the minimum difference with clinical significance. SEM and MDD were calculated as described by [21] after performing a repeated measurement ANOVA with the aov function in R. Due to the size of the sample (less than 50), normality was assessed with a Shapiro-Wilk test with the shapiro.test function in R. Homoscedasticity was assessed with a Levene test with the leveneTest function of the car package in R. Upper and lower limits of agreement (uLOA and lLOA) were calculated by conducting a Bland-Altman analysis with the blandr package in R.

To verify the validity of the algorithms, we calculated the correlation between their reported speed values and those obtained with the reference methods.

When appropriate, 95% CI are reported. The level of statistical significance was established at 0.05.

3 Results

A total of 18 people were recruited: 77.89 (6.47) y.o., 11 women. Valid signals were obtained for 15 of them: 78.53 (6.71) y.o, 9 women. GaitPy was able to detect walking activity at 2.4 m for 11 subjects only and for 14 of them at 6 m.

Table 1 and Table 2 show the results of the validity, agreement, and reliability analyses for short and long bouts, respectively.

Table 1. Results of validity, agreement, and reliability analyses for short bouts (2.4 m).

	Mueller's	GaitPy
Validity	r = 0.837 95% CI = (0.711, 1.000)	r = 0.724 95% CI = (0.492, 1.000)

Agreement	SEM = 0.082 m/s
	MDD = 0.228 m/s
	95% CI(uLOA) = (0.094, 0.319)
	95% CI(lLOA) = (−0.362, -0.137)
		SEM = 0.079 m/s
		MDD = 0.218 m/s
		95% CI(uLOA) = (0.066, 0.330)
		95% CI(lLOA) = (−0.370, -0.106)

Reliability	ICC(A, 1) = 0.812 95% CI = (0.532, 0.932)	ICC(A, 1) = 0.669 95% CI = (0.148, 0.899)

Table 2. Results of validity, agreement, and reliability analyses for long bouts (6 m).

	Mueller's	GaitPy
Validity	r = 0.799 95% CI = (0.649, 1.000)	r = 0.812 95% CI = (0.669, 1.000)

Agreement	SEM = 0.039 m/s
	MDD = 0.109 m/s
	95% CI(uLOA) = (0.042, 0.154)
	95% CI(lLOA) = (−0.176, −0.064)
		SEM = 0.072 m/s
		MDD = 0.199 m/s
		95% CI(uLOA) = (0.046, 0.251)
		95% CI(lLOA) = (−0.352, −0.147)

Reliability	ICC(A, 1) = 0.935 95% CI = (0.814, 0.978)	ICC(A, 1) = 0.799 95% CI = (0.476, 0.931)

4 Discussion

Validity assesses the ability of an algorithm's output to represent the target variable (in this case, gait speed). The point estimates for both Mueller's and GaitPy algorithms suggest that their validity values are good for both short and long bouts. However, their 95% CIs are too broad to support this conclusion. We cannot conclude that their validity is better than moderate; in fact, the lower end of GaitPy's CI for short bouts suggests that its validity could even be poor.

Agreement assesses the ability of an algorithm to provide the same value for multiple measurements on a stable subject. A difference of 0.1 m/s between two clinical walking tests taken two weeks apart is enough to trigger a geriatician's therapeutic response. Only Mueller's algorithm in long bout scenarios is a candidate to comply with such a requirement. It resulted in an SEM of 0.04 m/s and an MDD as low as 0.1 m/s, while the other three cases show an MDD twice as much. However, the widths of the 95% CIs of Bland-Altman's LOAs do not let us conclude that the agreement for Mueller's in long bouts scenarios is good enough to comply. Anyway, the 0.1 m/s threshold has been tested on measurement values from clinical tests. More studies are needed to test whether the same threshold applies to gait speed values estimated from free-living conditions. In fact, differences have already been observed between these two scenarios. For example, the values of gait speed, acceleration, and cadence in free-living conditions have been observed to be lower than those measured with clinical tests [10,13].

Reliability assesses the ability of an algorithm to provide different values for different subjects with different speeds. The point estimates for Mueller's algorithm suggest that its reliability is good for both short and long bouts. In particular, the 95% CI for long bouts let us conclude that reliability is, in fact, good or excellent. On the other hand, the 95% CI for short bouts does not let us conclude that its reliability is better than moderate. The point estimates for GaitPy suggest that its reliability is good for long bouts and moderate for short bouts. However, their 95% CIs suggest that its reliability could even be poor in both cases.

As expected, the overall performance of the algorithms is better for longer bouts than for shorter ones.

The main limitation of this preliminary study comes from its small sample size; which results in broad confidence intervals for the validity, agreement, and reliability estimators. We estimated that subsequent studies require sample sizes from 100 subjects on to be conclusive by running a simulation with synthetic data. We sequentially increased the sample size by adding duplicates of the data until the resulting 95% CIs were narrow enough.

5 Conclusion

Validity and reliability of state-of-the-art algorithms to estimate the gait speed of older adults in free-living conditions seem to be good. However, the results should be taken with caution because the small size of our sample results in broad confidence intervals for the estimators. The agreement seems not to be good enough to trigger therapeutic responses according to the current threshold for clinical tests. However, more studies are necessary to test whether the same threshold applies to gait speed estimations under free-living conditions.

Acknowledgements. This research is part of the ActiveUp project (grant number PID2019-108408RB-C21) funded by MICIN/AEI/10.13039/501100011033, the COVIDTECH-CM project funded by Comunidad de Madrid and the European Union via the European Fund for Regional Development as part of the Union's response to the COVID-19 pandemic, and was supported by CIBER - Consorcio Centro de Investigación Biomédica en Red - (CB06/01/0051), Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación.

References

1. Beard, J.R., et al.: The world report on ageing and health: a policy framework for healthy ageing. Lancet **387**, 2145–2154 (2016). https://doi.org/10.1016/S0140-6736(15)00516-4
2. Castro-Rodríguez, M., et al.: Frailty as a major factor in the increased risk of death and disability in older people with diabetes. J. Am. Med. Dir. Assoc. **17**, 949–955 (2016). https://doi.org/10.1016/j.jamda.2016.07.013
3. Chatterji, S., Byles, J., Cutler, D., Seeman, T., Verdes, E.: Health, functioning, and disability in older adults-present status and future implications. Lancet **385**, 563–575 (2015). https://doi.org/10.1016/S0140-6736(14)61462-8
4. Fried, L.P., et al.: Frailty in older adults: evidence for a phenotype. J. Gerontol. A Biol. Sci. Med. Sci. **56**, M146–M157 (2001). https://doi.org/10.1093/gerona/56.3.M146
5. Chen, X., Mao, G., Leng, S.X.: Frailty syndrome: an overview. Clin. Interv. Aging **9**, 433–441 (2014). https://doi.org/10.2147/CIA.S45300
6. Casas-Herrero, Á., et al.: Effects of Vivifrail multicomponent intervention on functional capacity: a multicentre, randomized controlled trial. J. Cachexia Sarcopenia Muscle **13**, 884–893 (2022). https://doi.org/10.1002/jcsm.12925
7. Alonso Bouzón, C., Carnicero, J.A., Turín, J.G., García-García, F.J., Esteban, A., Rodríguez-Mañas, L.: The standardization of frailty phenotype criteria improves its predictive ability: the toledo study for healthy aging. J. Am. Med. Dir. Assoc. **18**, 402–408 (2017). https://doi.org/10.1016/j.jamda.2016.11.003
8. Blinka, M.D., et al.: Developing a sensor-based mobile application for in-home frailty assessment: a qualitative study. BMC Geriatr. **21**, 101 (2021). https://doi.org/10.1186/s12877-021-02041-z
9. Galán-Mercant, A., Ortiz, A., Herrera-Viedma, E., Tomas, M.T., Fernandes, B., Moral-Munoz, J.A.: Assessing physical activity and functional fitness level using convolutional neural networks. Knowl.-Based Syst. **185**, 104939 (2019). https://doi.org/10.1016/j.knosys.2019.104939
10. Urbanek, J.K., Zipunnikov, V., Harris, T., Crainiceanu, C., Harezlak, J., Glynn, N.W.: Validation of gait characteristics extracted from raw accelerometry during walking against measures of physical function, mobility, fatigability, and fitness. J. Gerontol. A Biol. Sci. Med. Sci. **73**, 676–681 (2018). https://doi.org/10.1093/gerona/glx174
11. Cobo, A., et al.: Automatic and real-time computation of the 30-seconds chair-stand test without professional supervision for community-dwelling older adults. Sensors **20**, 5813 (2020). https://doi.org/10.3390/s20205813
12. Czech, M.D., et al.: Age and environment-related differences in gait in healthy adults using wearables. NPJ Digit. Med. **3**, 1–9 (2020). https://doi.org/10.1038/s41746-020-00334-y
13. Mueller, A., et al.: Continuous digital monitoring of walking speed in frail elderly patients: noninterventional validation study and longitudinal clinical trial. JMIR Mhealth Uhealth **7**, e15191 (2019). https://doi.org/10.2196/15191

14. Urbanek, J.K., et al.: Prediction of sustained harmonic walking in the free-living environment using raw accelerometry data. Physiol. Meas. 39, 02NT02 (2018). https://doi.org/10.1088/1361-6579/aaa74d
15. Kumar, D.P., Toosizadeh, N., Mohler, J., Ehsani, H., Mannier, C., Laksari, K.: Sensor-based characterization of daily walking: a new paradigm in pre-frailty/frailty assessment. BMC Geriatr. **20**, 164 (2020). https://doi.org/10.1186/s12877-020-01572-1
16. Gram-Hanssen, K., Darby, S.J.: "Home is where the smart is"? Evaluating smart home research and approaches against the concept of home. Energy Res. Soc. Sci. **37**, 94–101 (2018). https://doi.org/10.1016/j.erss.2017.09.037
17. Marikyan, D., Papagiannidis, S., Alamanos, E.: A systematic review of the smart home literature: a user perspective. Technol. Forecast. Soc. Chang. **138**, 139–154 (2019). https://doi.org/10.1016/j.techfore.2018.08.015
18. Czech, M.D., Patel, S.: GaitPy: an open-source python package for gait analysis using an accelerometer on the lower back. J. Open Sour. Softw. **4**, 1778 (2019). https://doi.org/10.21105/joss.01778
19. Ferre, X., et al.: Gait speed measurement for elderly patients with risk of frailty. Mob. Inf. Syst. **2017**, e1310345 (2017). https://doi.org/10.1155/2017/1310345
20. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2022). https://www.R-project.org/
21. Weir, J.P.: Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J. Strength Conditioning Res. **19**, 231–240 (2005). https://doi.org/10.1519/15184.1

Healthcare Based on IoT Using Arduino, MPU6050 Accelerometer and Gyroscope Sensor and FSR-174 Strain Gauge for Fatigue

Sara Fernández-Canosa[1], Verónica López González[1], Antonio Consoli[2], and Vanesa Soto-León[1](✉)

[1] FENNSI Group, National Hospital for Paraplegics, SESCAM, Toledo, Spain
`vani_vanesa@hotmail.com`
[2] ETSI de Telecomunicación, Universidad Rey Juan Carlos, Madrid, Spain

Abstract. The term "fatigue" refers to a change in task performance over time due to both psychological and physiological factors. No universal fatigue definition has been reached due to the strong subjective component attached to it. Fatigue assessment techniques vary from subjective scales to objective assessment tests such as isometric and finger tapping tasks performance. A low-cost, portable, and simple equipment is the most suitable option for the implementation of fatigue evaluation tasks during clinical visits.

The main goal of this work is to design and implement a biomedical device for muscle fatigue evaluation characterized by being portable, simple, and affordable. Additionally, correct functioning should be provided, and signal registration must be ensured.

For its development, an Arduino programmable board, a strain gauge sensor and a gyroscope and accelerometer sensor have been employed. The most appropriate sensors were selected: FSR-174 strain gauge and MPU6050 accelerometer and gyroscope sensor. Additionally, circuit design, assembly have been carefully implemented for the proposed goal.

Different fatigue measurements have been obtained and fatigue presence on the different recordings attributed by the new biomedical device have been demonstrated. As a conclusion, the design and implementation of an objective fatigue assessment equipment has been finalized and its correct functioning and signal registration capability have been proven. Near future daily clinic visits may provide the performance of fatigue assessment tasks with the biomedical electronic device created in this project after further investigations on register analysis, fatigue measurement computation and the device Internet connection.

Keywords: Fatigue · Finger-tapping task · Biomedical Electronic Device

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 138–148, 2023.
https://doi.org/10.1007/978-3-031-34586-9_11

1 Introduction

Fatigue is a condition commonly suffered among the general population. It can be defined as the maximal force decrease or difficulties in sustaining or initializing voluntary activities, however, there is no universal definition for fatigue since it is a subjective perception [1, 2].

Fatigue's high prevalence growth has given rise to a significant interest in the impact of fatigue on neurological disorders and neurorehabilitation [2]. When it comes to evaluating fatigue, there are no clear guidelines or a universal established standard leading to a wide variety of results. As a result, any posterior results comparison can be hardly carried out.

The measurement of fatigue by means of fatigue scales is a widespread technique. The scales that are usually used to evaluate the perceived fatigue are the Fatigue Severity Scale (FSS), the Modified Impact Fatigue Scale (MFIS) and the Borg scale [3, 4]. These scales rely on a strong subjective component which has given rise to the appearance of new different evaluation methods. This is the case of the Isometric (ISO) [5] and the Finger Tapping (FT) [6] performance tasks for fatigue assessment: two specific tasks that rely on an objective basis enhancing the posterior analysis of results. The ISO task entails sustained maximum voluntary contraction (MVC) over the time it is performed. Central fatigue is developed and motor units firing rate is diminished during maximal voluntary ISO tasks, leading to a decrease of voluntary activation [7]. Furthermore, literature reports high evidence of excitability reduction in the spinal cord and motor cortex produced by ISO tasks [8–10]. A smaller number of studies demonstrating the relationship between fatigue induced by means of repetitive movements have been reported [8, 10].

The FT test is a valid task to assess pathological and physiological mechanisms [11]. It is based on performing repetitive movements at the fastest possible rate leading to a frequency decrease in very few seconds from the beginning of the test. Frequency drop suggests fatigue induction.

ISO and FT tasks require a specific fatigue evaluation equipment which is usually expensive, bulky, and difficult to use [11, 12]. For these reasons, the following consequences are arisen:

- Expensive equipment: a reduced number of clinical health care groups and centers are provided with this particular equipment. It can be stated that it is an exclusive equipment.
- Bulky equipment: ISO and FT tasks can only be carried out to a limited number of subjects who can move to healthcare facilities equipped with the necessary biomedical devices for fatigue assessment. The possibility of carrying out fatigue evaluation tasks at subjects' homes is ruled out.
- Difficult to use: health professionals need training to be able to use this equipment and in most of the occasions it takes a long time to become familiar with it.

This is the case, for example, of the device we used in our previous studies [13, 14]: a general-purpose programmable data acquisition device (Biometrics DataLink DLK900 [15]) used with a goniometer and a dynamometer to perform ISO and FT tasks for fatigue assessment. There are other devices in the literature can use to FT and ISO, but the data obtain from these apparatuses cannot be used to measure the decrease in the motor performance [16, 17].

Due to the previous considerations, the main goal of this project is to develop a biomedical device for muscle fatigue assessment with the following characteristics:

- Valid and reliable: it should provide accurate fatigue measurements and be as effective as commercially available devices.
- Affordable: equipment cost should be reduced by means of low-cost devices (for example: open-source platforms, low-cost sensors…) making it possible to expand the use of fatigue assessment tasks to healthcare centers, outpatients, primary and secondary health care centers. In this way, fatigue assessment can be generalized.
- Portable: ISO and FT tasks will be carried out on a large scale, and it will be possible to perform fatigue evaluation tasks at subjects' homes.
- Simple: easy-to-use equipment should be developed to ensure no long learning procedures are required to healthcare operators.

The main goal of this study is to design and implement a biomedical device for muscle fatigue evaluation characterized by being reliable, affordable, portable and simple. Its correct functioning and valid signal registration must be ensured. Due to the presence of numerous studies that have demonstrated a strong relationship between both tasks (ISO task and FT task) and fatigue measurements these two tests are the basis of this project.

2 Device Design

In order to design and implement a low-cost equipment to assess fatigue, different devices and sensors have been chosen and evaluated. The device consists of an Arduino module and two sensors: a resistive strain gauge, for measuring the applied force, and a gyroscope and accelerometer module, for finger angle measurements.

2.1 Arduino

A commercially available microcontroller is used as the core of the device: Arduino UNO [18]. Arduino UNO is an open-source microcontroller-based board. It exposes 6 analog pins (input and output) and 14 digital pins (input and output). The Arduino Integrated Development Environment (IDE) can be used to develop and run the required software for muscle fatigue assessment test. Arduino UNO has been selected because of its portability, affordable price and flexibility in terms of input/output connections, of which only 5 will be used.

2.2 Sensors

A sensor to measure the force applied is required: it must have enough diameter to fit the index finger as this finger will be used to exert the force in the fatigue tests. For this reason, a resistive strain gauge, the Force Sensing Resistor FSR-174 from IEE Sensing, is used. For the detection of the index finger tilt angles, an Inertial Measurement Unit (IMU), consisting of a gyroscope and accelerometer module (TDK Invensense, MPU6050) is used. It is placed on top of the index finger during the fatigue test, taking advantage of its small size and light weight.

Strain Gauge: FSR-174

Strain is the deformation of a material when a stress is applied over it. A strain gauge is defined as an element with variable resistance which changes by means of tensile and compressive stresses. On the other hand, stress can be defined as the force exerted on a material divided by its cross-sectional area. In order to measure strain, the strain gauge must be connected to an electrical circuit capable of accurately detecting small resistance changes.

For the project development, a unique strain gauge connected to an electrical circuit has been used: the FSR-174. Its characteristics include: length, 63.7 mm; width, 27.8 mm; strain gauge resistance, higher than 1 MΩ; maximum electrical current, 1 mA; minimum and maximum operating temperature, −30 ºC, 170 ºC; force measurement, tension and compression stresses [19]. These characteristics have made it possible to measure resistance variations after surface pressure exertion thanks to a voltage divider.

Inertial Measurement Unit

The IMU combines an accelerometer and a gyroscope for the measurement of linear and angular accelerations, respectively.

The MPU6050 sensor is an IMU formed by 6 Degrees of Freedom (DoF): 3 axes to detect inertial forces (accelerometer) and 3 axes to detect rotations (gyroscope) to determine its instantaneous position. Its small size (21.2 mm × 16.4 mm × 3.3 mm) and weight (2.1 g) allows its use as a wearable sensor.

The accelerometer is responsible for detecting inertial forces applied to the sensor and projecting them onto three axes. The force direction decomposition into three axes of MPU6050 sensor is based in piezoelectric effect. The gyroscope is able to detect centrifugal forces and convert them into spin's velocity taking into account the three main reference axes: x, y, z. Combining the accelerometer and gyroscope measurements, it is possible to obtain the sensor's tilt angles and find its orientation [20].

3 Device Implementation

3.1 Hardware Connections

The circuit design, assembly and Arduino connections are described below. Circuit design is formed by 3 different main elements:

- Arduino UNO: a programmable microcontroller board made up of 6 analog inputs and 14 digital input/output pins.
- Strain gauge: it is formed by 2 pins, one of them connected to a 5 V source, the other one should be in series with A0 (analog pin) and a 10 kΩ resistance. The resistance should also be connected to the ground.

- IMU: 4 different connections are established between Arduino UNO and the sensor following the scheme below (see Fig. 1) [21], where A4 and A5 correspond to the analog ports 4 and 5 from Arduino UNO.

The complete circuit including all the aforementioned components and its positions can be identified (see Fig. 2).

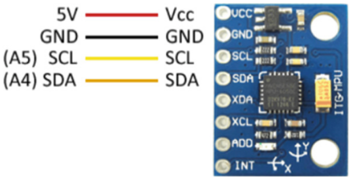


Fig. 1. Arduino UNO-MPU6050 connections.

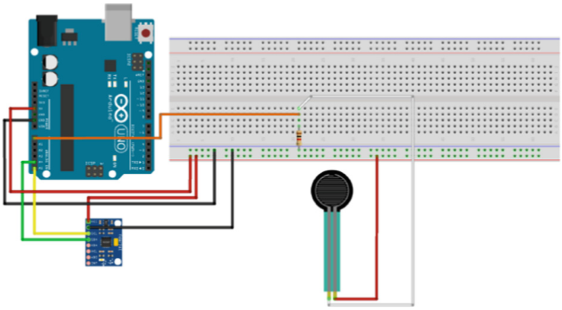


Fig. 2. Circuit design: components and connections.

3.2 Software Programming

The code running on the Arduino board was firstly divided in two different "sketches": strain gauge programming sketch and accelerometer and gyroscope sketch. After the completion of both sketches, they were unified in a single code.

A voltage divider is used to obtain the value of the resistance of the strain gauge, for converting the resistance reading into a voltage reading. To do this, the sensor is connected to an analogue pin that reads values between 0 and 1023. These values are converted to voltage (0–5 V) to obtain the strain gauge variable resistance.

In first place, it is necessary to add two libraries to control MPU6050 sensor, MPU6050.h and Wire.h. Subsequently the sensor initializes. In the next step, three variables storing raw values of inertial forces range from −2 g to 2 g (ax, ay, az) and three variables storing raw values of angular velocities range from −250°/second to 250°/second (gx, gy, gz). Moreover, the tilt angles of the sensor in the x and y axes are calculated.

3.3 Calibration Procedure

The IMU needs to be calibrated following a standard procedure [22]: The IMU is kept still on a flat surface and no movement should be performed. In this way, the expected or correct values should be ax = 0, ay = 0, az = 1 g for the acceleration among the 3 axes and gx = 0, gy = 0, gz = 0 for the rotation rates of the gyroscope. Offsets reading is continuously performed, and the values are corrected every 100 readings with the average offset. Finally, the corrected values are scaled to international units (*m*/*s*: ax = 0, ay = 0, az = 9.81; °/s: gx = 0, gy = 0, gz = 0).

Strain gauge calibration is defined as the procedure in which a weight value in kilograms (measured by the dynamometer of Biometrics Datalink) has been assigned to each voltage value (produced by FSR-174 strain gauge). It is carried out by means of 8 recordings in which the strain gauge has been superimposed on top of the dynamometer from Biometrics DataLink during data registration process for the voltage-weight value assignment. Each recording is composed of 25 s with the following structure: 5 s without applying force over the strain gauge, 15 s applying force continuously over the strain gauge at a specific voltage and 5 s without exerting force over the strain gauge.

4 Results

4.1 Data Acquisition Process and Fatigue Measurements

For the data acquisition process, ten different task recordings have been acquired: 5 FT task recordings and 5 ISO task recordings. All measurements have been acquired using the right index finger of the same and one-and-only subject, in order to obtain task recordings to compare the results provided from the electronic biomedical device that has been developed in this project and the current biomedical equipment Biometrics DataLink.

The IMU, comprising the accelerometer and gyroscope, is placed on the middle phalanx of the index finger. The strain gauge is placed just below the index finger in order to measure the pressure exerted by the index finger on it.

Furthermore, FSR-174 strain gauge is located on top of the dynamometer from Biometrics DataLink during data acquisition process for the posterior fatigue measurements comparison (see Fig. 3). Each FT recording and ISO recording stores five different data variables:

- Timestamp (ms): is the time in milliseconds from the starting point of the program to the end. Each data acquired during the recording is assigned a timestamp.
- Voltage (V): is the strain gauge voltage value that varies depending on the force applied towards it.
- Resistance (kΩ): is the resistance exerted by the strain gauge depending on the force applied towards it. The higher the force exertion, the lower the resistance value.
- X-angle (°): indicates the MPU6050's sensor tilt in the reference to the x-axis.
- Y-angle (°): indicates the MPU6050's sensor tilt in the reference to the y-axis. Data registration process involves different sampling rates for each biomedical device.

The sampling rate of the Arduino equipment is 85 Hz, approximately. On the other hand, Biometrics DataLink equipment sampling rate is 100 Hz. Once the data was

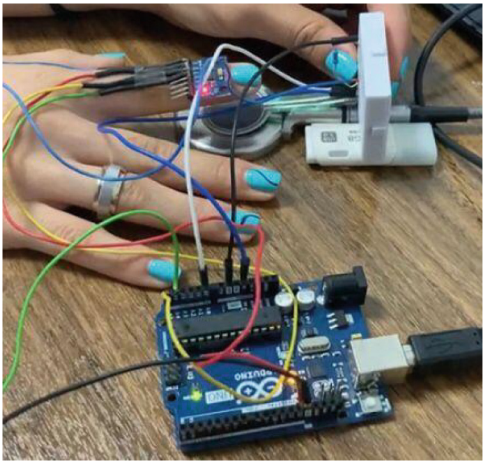


Fig. 3. Superposition of electronic equipment during data recordings for strain gauge calibration.

acquired with the biomedical device designed in this project, it was exported to.csv files by means of an Arduino plug-in named 'ArduSpreadsheet' for further analysis [22].

A unique fatigue measure for each task (ISO, FT) was evaluated and compared between the systems, the decay over the 2 min as a marker of the fatigability by computing the ratio of the motor output in the last 20 s compared with the first 20 s in both tasks [11, 13]. In the case of the ISO task, motor performance was measured as the ratio of the area of the curve between the initial and final 20-s blocks, while in the FT task it was measured as the ratio of the tapping frequencies. Areas were calculated using the so-called Q = trapz(Y) from MATLAB and the number of finger taps was calculated using the so-called [pks,locs] = findpeaks (data) function from MATLAB.

Strong differences in the acquired signal during the same registration process can be seen during ISO task and FT task (see Fig. 4 and Fig. 5). With respect to similarities during the ISO task, in all the recordings regardless of the equipment, the area obtained in the first 20 s is higher than the area obtained in the last 20 s. As a consequence, muscle fatigue is present in all the recordings: force exertion declines over time due to muscle fatigue in the index finger which performs the MVC.

Regarding FT task, as it can be observed in Fig. 5, the measurement of the tapping frequency obtained by means of peak detection is more correct with the Arduino device. The algorithm is not capable of marking the peak extreme of some finger taps recorded with Biometrics DataLink equipment. Moreover, it should be noted that the ratio of the tapping frequency between the first 20 s and the last 20 s of the task, as measured by the Arduino board, is greater than 1, which indicates fatigue.

5 Discussion

There is a need to design and implement a new biomedical device for muscle fatigue assessment because the currently available commercial equipment do not meet the needs for portability and low costs of healthcare.

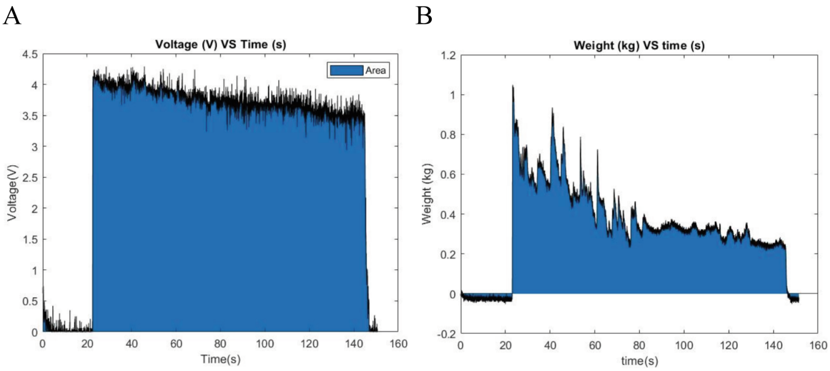


Fig. 4. (A) Area of the signal acquired with Arduino (B) Area of the signal acquired with Biometrics DataLink device during ISO task.

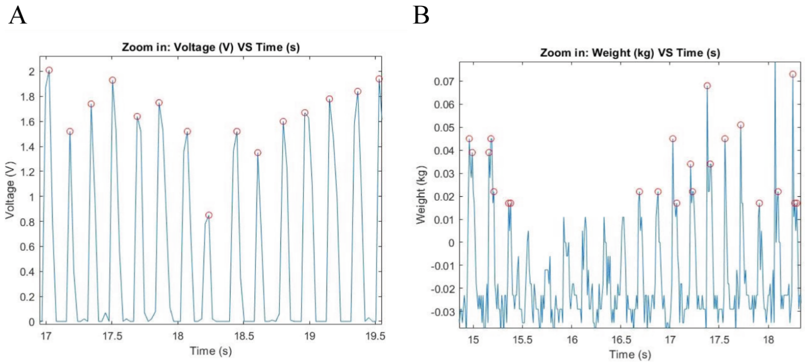


Fig. 5. (A) Zoom in of FT task signal with local maxima for Arduino equipment (B) Zoom in of FT task signal with local maxima marks for Biometrics equipment.

Different features take part in fatigue assessment equipment: complexity, static devices, non-portable equipment and high-cost. Each of these characteristics have been challenged thanks to the design and development of the biomedical device carried out in this project for muscle fatigue evaluation.

The following attributes characterize this new equipment with the same purpose, fatigue assessment:

- Simplicity: it has been accomplished by means of open-source platforms for electronic projects like Arduino. Moreover, plain sensors, which have been previously used by a large number of users worldwide, have been manipulated: strain gauge sensor and accelerometer and gyroscope sensor.
- Portable: Arduino-computer connection is established by means of a USB cable. In this way, the biomedical equipment and a computer are the two necessary elements for the development of fatigue evaluation tasks. As a consequence, ISO and FT tasks could be carried out in subjects' homes for populations who cannot attend healthcare facilities.

- Low-cost: the biomedical device equipment cost is established at 40€, approximately. For this reason, fatigue assessment equipment is characterized by being affordable falling within the healthcare budget.

Apart from the design and implementation of the new biomedical device, reliable recordings derived from its correct functioning have been achieved, capable of measuring fatigue induced in two 2-min fatigue tasks performed with the index finger, an isometric task and a finger tapping task. In this way, the main goal of the project has been completed: the creation of a biomedical device characterized by its simplicity, portability and low-cost that ensures solid and valid fatigue registrations thanks to its correct functioning.

In relation to fatigue measurements calculations, further analysis should be performed. The acquisition method for ISO and FT signals, in which equipment overlapping has taken place, has given rise to imprecise signal acquisitions with Biometrics DataLink equipment. The equipment of Biometrics DataLink is designed for direct signal acquisition. As a consequence, erroneous fatigue measurements have been obtained: frequency ratio and area ratio for FT and ISO tasks, respectively.

A frequency analysis (for example, Fourier transform analysis) for signal recordings may provide a more valid and reliable fatigue measurement than the peak detection algorithm implemented in this project.

A higher number of registrations should be acquired for solid fatigue measurements. An average of all of them should be performed in order to establish a valid and reliable equipment comparison.

5.1 Limitations

Different limitations have arisen in the course of this project which have affected its development and the outcomes of the new biomedical device:

- Strain gauge diameter has constrained calibration procedure by means of heavy and small weights. As a result, only few voltage-weight estimations accomplished through force exertion recordings have been obtained. The low number of registrations used for strain gauge calibration constrained these approximations.
- The superimposition of both equipment for posterior fatigue measurements analysis have given rise to imprecise Biometrics DataLink fatigue signals due to the essence of this equipment: nothing but the index finger should be on top of it during signal acquisition.
- Peak detection algorithm for FT tasks have demonstrated imprecise results for Biometrics DataLink acquisition. In this way, erroneous ratios have been obtained for an already tested and valid fatigue assessment equipment.
- A greater number of registrations entails more accurate and therefore, well-founded fatigue measurements attainment.
- We tested the device only on one subject, more tests should be done with real patients.

6 Conclusions

The result of this project shows how a complex, static and high-cost equipment the use of which is restricted to a small group of population due to its characteristics; can inspire the design of a simple, portable, and low-cost equipment the use of which can be extrapolated to daily clinical visits. Its correct functioning and solid data acquisition can provide healthcare professionals objective tests for fatigue assessment that can be standardized in the near future.

Further research should follow analysis of the recorded data and strain gauge calibration. Potential improvements for this electronic device include transmission of the data will to a server and the possibility of wireless connection to the Internet.

Funding. This research was funded by the Fundación Ramón Areces, XX Concurso Nacional para la adjudicación de Ayudas a la Investigación en Ciencias de la Vida y de la Materia.

References

1. Chen, M.K.: The epidemiology of self-perceived fatigue among adults. Prev. Med. **15**(1), 74–81 (1986)
2. Chaudhuri, A., Behan, P.O.: Fatigue in neurological disorders. Lancet **20**(363), 978–988 (2004)
3. Anton, H.A., Miller, W.C., Townson, A.F., Imam, B., Silverberg, N., Forwell, S.: The course of fatigue after acute spinal cord injury. Spinal Cord **55**(1), 94–97 (2017)
4. Mordillo-Mateos, L., et al.: Fatigue in multiple sclerosis: general and perceived fatigue does not depend on corticospinal tract dysfunction. Front. Neurol. **9**(10), 339 (2019)
5. Al-Mulla, M.R., Sepulveda, F., Colley, M.: A review of non-invasive techniques to detect and predict localised muscle fatigue. Sensors **11**(4), 3545–3594 (2011)
6. Shimoyama, I., Ninchoji, T., Uemura, K.: The finger-tapping test: a quantitative analysis. Arch. Neurol. **47**, 681–684 (1990)
7. Cudeiro-Blanco, J., et al.: Prevalence of fatigue and associated factors in a spinal cord injury population: data from an internet-based and face-to-face surveys. J. Neurotrauma **34**(15), 2335–2341 (2017)
8. Gandevia, S.C.: Spinal and supraspinal factors in human muscle fatigue. Physiol. Rev. **81**(4), 1725–1789 (2001)
9. Butler, J.E., Taylor, J.L., Gandevia, S.C.: Responses of human motoneurons to corticospinal stimulation during maximal voluntary contractions and ischemia. J. Neurosci. **23**(32), 10224–10230 (2003)
10. Di Lazzaro, V., et al.: Direct demonstration of reduction of the output of the human motor cortex induced by a fatiguing muscle contraction. Exp. Brain Res. **149**(4), 535–538 (2003)
11. Arias, P., et al.: Central fatigue induced by short-lasting finger tapping and isometric tasks: a study of silent periods evoked at spinal and supraspinal levels. Neuroscience **305**, 316–327 (2015)
12. Mordillo-Mateos, L., et al.: Fatigue in multiple sclerosis: general and perceived fatigue does not depend on corticospinal tract dysfunction. Front. Neurol. **10**, 339 (2019)
13. Onate-Figuérez, A., et al.: Hand motor fatigability induced by a simple isometric task in spinal cord injury. J. Clin. Med. **11**(17), 5108 (2022)
14. Soto-Leon, V., et al.: Effects of fatigue induced by repetitive movements and isometric tasks on reaction time. Hum. Mov. Sci. **73**, 102679 (2020)
15. Specialist manufacturer of equipment for data acquisition research and clinical rehabilitation. http://www.biometricsltd.com. Accessed May 2022
16. Boukhvalova, A.K., et al.: Identifying and quantifying neurological disability via smartphone. Front. Neurol. **9**, 740 (2018)
17. Janković, M.M., Popović, D.B.: An EMG system for studying motor control strategies and fatigue. In: 10th Symposium on Neural Network Applications in Electrical Engineering (2010)

18. What is Arduino? http://www.arduino.cc. Accessed Apr 2022
19. SS-U-N-S-00039 | Galga extensiométrica de perfil bajo I.E.E., >1 MΩ | RS Components. (s. f.). RS-online. https://es.rs-online.com/web/p/galgas-extensiometricas/1895590 2022/04. Accessed May 2022
20. Los sistemas de medida inercial. Tienda y Tutoriales Arduino. https://www.prometec.net/imu-mpu6050/. Accessed 05 Mar 2022
21. Determinar la orientación con Arduino y el IMU MPU-6050. https://www.luisllamas.es/arduino-orientacion-imu-mpu-6050/. Accessed 08 Mar 2022
22. Logging Arduino Serial Output to CSV/Excel (Windows/Mac/Linux) - Circuit Journal. Circuitjournal.com. https://circuitjournal.com/arduino-serial-to-spreadsheet. Accessed June 2022

Smartphone-Based Strategy for Quality-of-Life Monitoring in Head and Neck Cancer Survivors

Laura Lopez-Perez[1](✉), Itziar Alonso[1], Elizabeth Filippodou[2],
Franco Mercalli[3], Stefano Cavalieri[4,5], Elena Martinelli[5], Lisa Licitra[4,5],
Anastassios Manos[2], María Fernanda Cabrera-Umpierrez[1],
and Giuseppe Fico[1]

[1] Universidad Politécnica de Madrid-Life Supporting Technologies Research Group, ETSIT, Madrid, Spain
llopez@lst.tfo.upm.es
[2] Information Technology Programme Management Office, DOTSOFT, Thessaloniki, Greece
[3] MultiMed Engineers srls, Parma, Italy
[4] Head and Neck Medical Oncology Department, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[5] Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy

Abstract. Smartphone-based IoT systems have the potential to predict and keep control of Quality of Life measurements with the adequate design. This work aims to provide a comprehensive strategy for the development of a disease-specific application to monitor Quality of Life in Head and Neck Cancer survivors. This research first presents the results from a literature review focused on mHealth services for cancer patients. Second, we provide a complete overview of a clinical trial protocol where patients are encouraged to (1) perform self-management by actively reporting symptoms, (2) keep healthy lifestyles, (3) interact with an embedded artificial intelligence that serves as an additional patient-physician communication channel, and (4) fill in Quality of Life standard questionnaires from a web platform from home. The challenges addressed, the unobtrusive data collection procedures chosen, and the quality data obtained from physical, social, and behavioral measures, provide a resourceful set of guidelines and requirements for future research works aimed at after-treatment cancer patients monitoring through IoT portable devices.

Keywords: IoT systems · IoT-based monitoring systems · cancer research · head and neck cancer · quality of life

1 Introduction

The advances in digital technologies for data gathering, analysis and monitoring of patients has allowed the incorporation and standardization of novel healthcare services, such as Internet of Things (IoT) systems, that constitute portable cloud-connected tools able to record personal metrics unobtrusively, through sensors and custom applications.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 149–158, 2023.
https://doi.org/10.1007/978-3-031-34586-9_12

The most prominent IoT devices are smartphones, mobile devices integrated in daily life routines. Their huge potential for health monitoring and disease management is being widely explored by researchers and physicians. In particular, eHealth smartphone applications are currently being used for patient self-management, continuous symptom monitoring in chronic diseases, and as an additional communication bridge between patients and physicians [1, 2]. The engagement of individuals with smartphone applications that promote healthier and stress-free lifestyles have shown general positive outcomes and, precisely, cancer patients showed high technology acceptance [3], which is a major reason why the use of IoT in cancer research is growing. In terms of diagnosis, for instance, work has been done to use IoT (i.e., tactile neuron devices) with Artificial Intelligence (AI) technology for more accurate cancer diagnosis [4]. Other IoT devices are commonly embedded or synchronized with smartphones (e.g., sensors and wearables), gathering a broader diversity of patient data. The continuously updating of technologies has led to the integration also of 5G for faster communication, and blockchain services to ensure the security of the data gathered [5].

Current data-driven strategies aid clinicians on continuous monitoring the quality of life of patients and serve to raise public awareness of the importance of prevention measures in modifiable risk factors for cancer. Physical and psychological problems can be derived because of treatment causing long-term consequences. Therefore, preventing long-term effects, and anticipating the early deterioration of quality of life is of paramount importance to minimize the effect of these problems in their daily routines [6]. Apart from collecting patient self-reported outcomes, through standard questionnaires or other analyzed surveys, multiple domains of well-being using big data can offer valuable information for monitoring health status after cancer treatment. Information collected from traditional sources of health data (e.g., electronic health records (EHR) or clinical registries) being complemented with new sources of data (i.e., IoT such as mobile health) could support this continuous monitoring and therefore prevent Quality of Life (QoL) deterioration [7].

Head and Neck Cancers (HNC) are a malignant neoplasm group that mainly develops in the squamous cells that line the mucosal surfaces inside the head and neck [8]. In 2020, this cancer may have affected approximately 151,000 new patients in Europe and 833,000 new patients worldwide [9], being the seventh most common cancer worldwide. In the last five years, the cancer survival rate has increased due to a better knowledge of the scientific advances on the origin of cancer and the existence of new treatments that are improving day by day [10]. This rise of cancer survival rates has led to a higher number of survivors, increasing the need to emphasize health-related quality of life (HRQoL). As a result, the study of HRQoL has grown significantly and has become an essential component in cancer care in the last years [11]. Among all the HNC survivors, their QoL is mainly affected by the pain perceived in the cervical and shoulder regions, the perception of their physical fitness, and the long-term fatigue reported after completion of medical treatment. Therefore, an adequate treatment strategy is needed so that the quality of life of these survivors is not diminished. For this reason, the use of IoT can facilitate and obtain a better treatment strategy and help maintain and improve the QoL of HNC survivors, through the use of health-related data and patient monitoring [12]. To ultimately improve the QoL of Head and Neck Cancer (HNC) survivors, IoT tools can help on controlling adverse effects, signs and symptoms, physical activity routines, and sleep patterns [13].

The use of a mobile application for patient monitoring, as well as the use of their data, entails several challenges, such as data quality, transparency, data protection and trustability. Current personal data behavior monitoring based on smartphones used by patients, include two main categories of data collected: a) raw data from smartphone sensors, or b) calculated, more advanced level of data from third party applications (e.g., Samsung, Google). The decision on which of these two types of data should be collected is based on privacy, unobstructiveness, and clean (or low energy) computing on the edge. Regarding the data protection, privacy issues emerge when collecting data from third parties' data hubs, which need to be addressed through encryption and/or pseudoanonymisation techniques. If only raw data collected and if data analysis processes are implemented only at the edge node, then privacy can be ensured at highest level. Although raw and some calculated data can be collected unobstructively, there are types of personal IoT based data (such as sleep behavior or skin temperature) that can be currently only collected through the user's specific action. Finally, major challenge remains the ability to enable as low and clean computing energy at the edge (the smartphone device), to allow resources (battery, memory) to remain at an adequate level for both data collection and analysis, without blocking the rest of the smartphone operations.

There are also other types of challenges when developing health monitoring services, these are the difficulties related to patients. The receptiveness of patients to use digital technology after-treatment is closely related to their motivation. The degree of willingness to use tools or technologies to monitor physical activity depends on the patient's attitude towards physical activity, and the patient's attitude toward technology-assisted physical activity. These two characteristics reflect the motivation of cancer survivors to exercise and use technology, since not all patients are receptive to and able to use digital technologies to improve their physical activity [14, 15]. Another difficulty is the level of technology literacy of cancer survivors, a high level of use of technologies (e.g., computers, smartphones, internet etc.), has a positive impact in the level of physical activity and QoL of the participants [16]. In addition, another of the great barriers is the possibility to have internet access, especially in people over 65 years of age. The cancer survivors that lack digital services that work and suit individual needs, are more likely to have a decrease in motivation and use of eHealth technology, which can lead to a decrease in their physical activity and a worsening of their QoL [15].

To solve these challenges, the BD4QoL project aroused, whose main objective is to improve HNC survivor's Quality of Life through person-centered monitoring and follow-up planning by contribution of artificial intelligence (AI) and big data (multidisciplinary medical, environmental, personal feelings, socioeconomic and behavioral data) unobtrusively collected from commonly used mobile devices, in combination with multi-source clinical, socioeconomic data and patients reported outcomes, to profile HNC survivors for improving personalized monitoring and support. The analysis of newly QoL indicators will allow anticipating risks, inform patients and caregivers for personalized interventions to timely intercept and prevent long-term treatment effects.

2 Materials and Methods

This project is based on a multidisciplinary collaboration. The BD4QoL Consortium is a strong and balanced partnership composed of 16 organizations, located in 7 different European countries, including: 3 outstanding cancer reference centers, 6 research institutions, 2 large corporation and 1 industrial partner, 2 SMEs and 2 Government/Public bodies [17]. With this work modality, the aim is to take the maximum advantage of individual expertise form each team of experts to monitor the clinical research and the technology challenges which the project brings about. Figure 1 represents the steps that have been followed to carry out this study: a literature review, analysis of the possible scenarios, implementation of the mobile app and services, and the ongoing validation in a clinical trial.

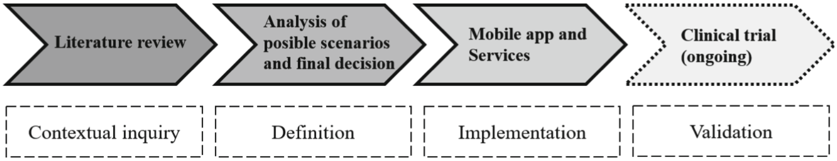


Fig. 1. Diagram of the materials and methods of the study.

To assess which mHealth devices have being used for QoL monitoring, a literature review was conducted. It was carried out following search terms “mhealth app for QoL monitoring” and “wearable devices in cancer research” in PubMed in February 2021. Three different authors search first for title and abstract and later of the full-text manuscript and reject the ones that were not satisfying the scope (i.e., studies using mobile apps to monitor QoL). The goal of this search was to identify conventional operating systems, most convenient measures from smart devices, minimum data required for daily monitoring, challenges addressed, and quality rules considered in the implementation. Limitations were also analyzed and filtered for the case study of cancer survivors. With the results of the literature review, we evaluated and proposed to clinical partners a set of scenarios with pros and cons to evaluate, discuss and agree together with is the most appropriate one to implement within the BD4QoL project.

A prospective study (randomized controlled trial) has been set-up and launched where HNC survivors are enrolled and randomly assigned to an intervention arm equipped with mobile apps for continuous monitoring and support or to a control arm that will be monitored by means of QoL standard questionnaires issued by the European Organization for Research and Treatment of Cancer (EORTC). More than 400 HNC survivors are expected to be enrolled in this study, coming from four differences centers in Italy and UK. Details of the protocol can be found at ClinicalTrials.gov with the NCT05315570 identifier. This is a multicenter, international, two-arm, randomized (2:1 ratio), open label, superiority trial, designed to evaluate the proportion of HNC survivors experiencing a clinically meaningful QoL deterioration (reduction of at least 10 points in EORTC QLQ-C30 global health status [18, 19]) between at least 2 visits during post-treatment follow-up (up to 24 months from randomization) with the use of BD4QoL platform in comparison to those without the BD4QoL platform ("standard FU"). Mobile data from intervention arm participants will be also used to recognize behavioral changes due to the participants' mobile usage. In addition, BD4QoL project aims to assess other mobile app features like lifestyle self-management, counseling, communication and detection of affective traits. With the mobile apps and QoL proxies' data collection, risk models will be implemented to possibly identify behavioral changes that might be associated with QoL modifications (improvement or deterioration) in association with patient's reported outcomes and experience measures (PROMs/PREMs). This clinical trial has been carried out in conjunction with the clinical and technical partners of the project to include all the relevant aspects, using the SPIRIT AI [20] and CTTI [21] guidelines for clinical trials.

3 Results

3.1 Literature Review

As a result of the literature review, we collected and examined 67 papers, from peer-reviewed journals and conference proceedings, illustrating the application of IoT and mobile technologies in healthcare research. The main characteristics extracted from these papers are summarized in Table 1.

Challenges addressed cover 1) telemonitoring, to enable healthcare professionals (including practitioners and nurses) to care for more patients with no decrease in quality of service; 2) self-monitoring and symptoms self-management, to empower patients to deal with the less risky aspects of their health and well-being, reducing the need for resorting to more expensive healthcare resources; 3) counselling, to coach patients towards healthy behaviors that will prevent future health decays; 4) collection of PROM and PREM questionnaires, to assess quality of life and quality of care dimensions; and 5) support for adherence to therapy and medications.

The wide number of clinical domains covered, also summarized in Table 1, is testament to how IoT and mobile health technologies promise to transform the whole healthcare landscape in the coming years. Applications are directed to patients suffering from (possibly multiple) chronic conditions, who need long term care; patients recovering from surgery who need continuous support, although for a limited, pre-defined period; patients with specific diseases, that each have their correspondingly specific requirements in terms of remote support to be provided, etc. Healthy subjects are also targeted with health promotion applications, addressing primary prevention.

On the interoperability side, 50% of papers report the specific operating system they relied on. The majority of research endeavors still rely on the Android operating system, likely due to its higher *openness*, which is mentioned in 49% of the papers. However, 24% of the papers also support iOS devices, in addition to Android. Only 1 paper report supporting iOS only.

Information on data collection and data quality is still not widespread. Only 9% of the examined papers report explicitly the strategy for collecting the data they used (e.g., frequency of collected measurements, procedures to be enacted by users, etc.). On the other hand, a relevant number of papers (19%) reports about data quality rules. In many cases, these are in the form of post-hoc rules included in study protocols to determine

Table 1. Characteristics of surveyed papers.

Total number of papers reviewed	n = 67
Challenges addressed *Telemonitoring, self-monitoring, symptoms self-management, counselling, PROM/PREM collection, adherence promotion*	
Clinical domains addressed *Health promotion (physical activity, nutrition, diet), asthma, cancer, obesity, heart failure, stroke, peripheral arterial disease, diabetes, Parkinson disease, kidney disease, inflammatory bowel disease, amyotrophic lateral sclerosis, chronic obstructive pulmonary disease, complex chronic conditions, orthopedy, surgery follow up, pain treatment, autism, mental health*	
Papers reporting mobile OS	25% (n = 17)
Only Android	1% (n = 1)
Only iOS	24% (n = 16)
Both	
Papers reporting information on data collection protocol	9% (n = 6)
Papers reporting information on data quality rules	*19% (n = 13)*

when a participant is to be part of the analysis. However, in other cases, quality rules to be specifically enacted during the data collection process (e.g., minimum percentage of data collected by the step counter over daily hours, for the data point to be considered as valid), are indicated. This is very important, for instance, in those cases in which the IoT data are used to enact a clinical workflow, such as for instance when alerting the Point of Care that a check with the patient is recommended, to verify why her physical activity measurement (e.g., obtained through the pedometer) consistently decreased over the last few days.

3.2 Analysis of Possible Scenarios and Final Decision

Data collection of behavioral (personal) being accessible and available for analysis data requires to be authorized under a secure network and GDPR compliant procedures. The main challenge here is to allow data collection of all three agreed BD4QoL domains of analysis, named a) physical, b) social and c) sleep behavioral of patients, while satisfying at the same time the need of obstructiveness and protocol compliance.

Towards this direction, the main options identified involved to a) include or not include personal wearable devices, named smartwatches, and b) to allow use of both Android and iOS smartphone operating systems used by patients, under the intention to include as much as possible clinical trial population. These options were analyzed with pros and cons for each followed scenario, as described below.

a) Use of smartwear devices – smartwatches:
   a. Pros
      - Accurate sleep monitoring behavior.

  - Accuracy in physical behavior.
  - Data collection can include a more *mature* and coherent set of measurements based on algorithms using raw data sensors.
  - Low level granularity of data measurements (data can be collected at different levels of granularity, from less than 1 s to summary aggregated daily data through third party web services).
  - Low energy utilization of resources based on smartwatch data.

b. Cons

  - Technical implementation is device specific unless a completely different architecture is followed that could lead to scalability.
  - Obstructiveness of clinical experiment which could lead to faulty assumptions based on mis-wearing behavior of wearables.
  - Cost is high for both use of wearables and API access for some brands.
  - High energy resources for batter and memory capacity needed to compute data collection and baseline calculations at the edge level, i.e., the smartphone device.

b) Use of Android compared to iOS for smartphone devices:

a. Pros

  - Ability to include a larger number of populations for the clinical trial.
  - Greater level of un-objectiveness of the experiment.

b. Cons

  - Missing the social domain related data for the clinical experiment since iOS operating systems do not allow related data to be collected for privacy issues (e.g., Call and SMS logs).
  - Data collection of location related activities could be available (at the time of design decisions) every 15 min for iOS devices, in contrast with 1 min data measurements collected for Android devices, concluding in several implications for the recognition algorithm used to recognize the semantic location of the places one visits.

Based on the above analysis, it was decided to opt for the use of only Android smartphones for data collection (Fig. 2), based on a mixed scenario for the acquisition process. Raw level data such as screen status (ON, OFF), light level, accelerator, and GPS measurements would be collected through the phone's baseline sensors, whereas more higher-level data groups, such as steps, physical activities and phone based social data (i.e., calls, SMS, applications used, etc.) would be collected through third party API services (Google Fit and device OS packages). To enable compliance with GDPR, all data are pseudoanonymized, whereas sensitive private social data (such as calls and SMS) are also encrypted.

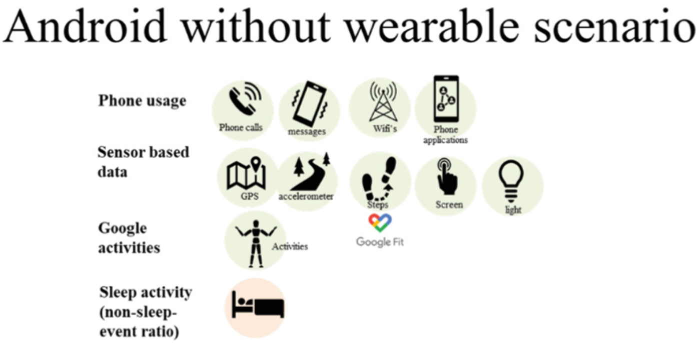


Fig. 2. Features included in the scenario selected

3.3 Implementation of the Mobile App and Services

The BD4QoL platform obtained as a result, is made up of a mobile application that collects patient data, and a Point of Care tool where clinicians view and manage patient data. The main objective of the mobile app is to continuously collect data from the device's available sensors and operating system. It also has self-administered questionnaires, for the collection of additional QoL items, to be measured between one follow up visit and the next, including supporting sub-functions (e.g., reminders, consistency checks). The mobile app also includes a self-management e-coach, consisting of a patient empowerment, natural language-based chatbot, offering to participants: 1) visualization of their own QoL data and related trends, as collected and inferred by the platform, according to rules established with clinicians, 2) counseling on symptoms self-management and providing relevant suggestions and recommendations for guided self-help, 3) establishing a communication channel with the clinician, and 4) detecting affective traits relevant to the participant's QoL assessment (e.g. depression, anxiety) from the analysis of the dialog among the participant and the chatbot, through natural language processing and understanding algorithms.

Finally, machine learning-based data analysis algorithms that use the data collected through the mobile app during the prospective BD4QoL trial, will be developed to improve predictive modeling for the early detection of HRQoL or other health outcome deterioration.

4 Discussions and Conclusions

The concept of IoT is expanding rapidly and has been integrated into the standard of living for a great portion of the population in the form of smart devices, such as personal mobile phones. Researchers and physicians are taking advantage of this technology's readiness to enhance healthcare workflows, but every health condition has its particular needs, and thus disease-specific settings must be designed to optimize the performance of the application. In this work we have presented the results obtained from a literature review in terms of mHealth applications for QoL monitoring and the use of wearable devices in cancer research, demonstrating that IoT is a fast-growing and broadly applicable tool for improving conventional healthcare strategies applied in the management of cancer survivors. Additionally, the design of the app ensured fundamental principles to avoid any patient's discomfort: unobtrusiveness guarantees that the system does not interfere with the patient's routines unless willing, high security measures provide confidence that no personal data will be breached, privacy rights are preserved during the monitoring process and, to provide trust, the application follows transparency criterions by allowing users to check which kind of data is collected, permissions can be removed and withdrawal is permitted at any time with no detrimental to the patient's standard of care protocol.

Previous works already noticed the role of smartphones as pervasive tools to entangle clinical procedures with IoT-based monitoring systems for other cancer patients [22, 23] but a specific design for Head and Neck cancer patients was missing. More importantly, the needs of patients going through treatment differ from survivors dealing with after-treatment sequelae, a chronic condition that is usually revised only though scheduled but sporadic medical visits.

Self-management of the disease, patient empowerment and promotion of healthy lifestyles are clear strengths of IoT-based strategies for cancer survivorship. However, there is still room for improvement in terms of machine learning algorithms, not analyzed in this work, to optimally gather clinically relevant knowledge from IoT devices and offer the best possible support to these patients. With the emergence of an avalanche of healthcare applications in the market, the use of data model standards to ensure semantic harmonization and interoperability will provide an unprecedented multi-source resource to better understand disease progression and personalize out-of-the-clinic therapies.

Thus, the BD4QoL application is a co-creation effort from a multidisciplinary team and the guidelines obtained from this work will serve future studies and trials to develop tailored disease specific IoT applications for better monitoring of quality of life.

References

1. Hamine, S., Gerth-Guyette, E., Faulx, D., Green, B.B., Ginsburg, A.S.: Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review. J. Med. Internet Res. **17**(2), e52 (2015). https://doi.org/10.2196/jmir.3951
2. Kwan, Y.H., et al.: Evaluation of mobile apps targeted at patients with spondyloarthritis for disease monitoring: systematic app search. JMIR mHealth uHealth **7**(10), e14753 (2019). https://doi.org/10.2196/14753
3. Moses, J.C., Adibi, S., Shariful Islam, S.M., Wickramasinghe, N., Nguyen, L.: Application of smartphone technologies in disease monitoring: a systematic review. In Healthcare, vol. 9, no. 7, p. 889. MDPI (2021). https://doi.org/10.3390/healthcare9070889
4. Evans, S.: Researches use IoT for cancer diagnosis (2022). https://www.iotworldtoday.com/2022/07/14/researches-use-iot-for-accurate-cancer-diagnosis/. Accessed 18 Oct 2022
5. Muhsen, I.N., et al.: Current status and future perspectives on the Internet of Things in oncology. Hematol. Oncol. Stem Cell Ther. (2021). https://doi.org/10.1016/j.hemonc.2021.09.003
6. Alonso, I., Lopez-Perez, L., Guirado, J.C.M., Cabrera-Umpierrez, M.F., Arredondo, M.T., Fico, G.: Data analytics for predicting quality of life changes in head and neck cancer survivors: a scoping review. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2262–2265. IEEE (2021)

7. FUNDING & TENDERS: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/sc1-dth-01-2019. Accessed 18 Oct 2022
8. Head and Neck Cancer—Patient Version (2022). https://www.cancer.gov/types/head-and-neck. Accessed 18 Oct 2022
9. Boscolo-Rizzo, P., et al.: The evolution of the epidemiological landscape of head and neck cancer in Italy: is there evidence for an increase in the incidence of potentially HPV-related carcinomas? PLoS ONE **13**(2), e0192621 (2018). https://doi.org/10.1371/journal.pone.0192621
10. Kocarnik, J.M., et al.: Cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life years for 29 cancer groups from 2010 to 2019: a systematic analysis for the global burden of disease study 2019. JAMA Oncol. **8**(3), 420–444 (2022). https://doi.org/10.1001/jamaoncol.2021.6987
11. Taylor, K., DipOnc, G., Monterosso, L.: Survivorship care plans and treatment summaries in adult patients with hematologic cancer: an integrative literature review. In: Oncology Nursing Forum, vol. 42, no. 3 (2015). https://doi.org/10.1188/15.ONF.283-291
12. Ortiz-Comino, L., et al.: Factors influencing quality of life in survivors of head and neck cancer: a preliminary study. In: Seminars in Oncology Nursing, vol. 38, no. 4, p. 151256. WB Saunders (2022). https://doi.org/10.1016/j.soncn.2022.151256
13. de Queiroz, D.A., da Costa, C.A., de Queiroz, E.A.I.F., da Silveira, E.F., da Rosa Righi, R.: Internet of things in active cancer treatment: a systematic review. J. Biomed. Inform. **118**, 103814 (2021). https://doi.org/10.1016/j.jbi.2021.103814
14. Rossen, S., Kayser, L., Vibe-Petersen, J., Christensen, J.F., Ried-Larsen, M.: Cancer survivors' receptiveness to digital technology–supported physical rehabilitation and the implications for design: qualitative study. J. Med. Internet Res. **22**(8), e15335 (2020). https://doi.org/10.2196/15335
15. Duman-Lubberding, S., van Uden-Kraan, C.F., Peek, N., Cuijpers, P., Leemans, C.R., Verdonck-de Leeuw, I.M.: An eHealth application in head and neck cancer survivorship care: health care professionals' perspectives. J. Med. Internet Res. **17**(10), e4870 (2015). https://doi.org/10.2196/jmir.4870
16. Ester, M., McNeely, M.L., McDonough, M.H., Culos-Reed, S.N.: A survey of technology literacy and use in cancer survivors from the Alberta Cancer Exercise program. Digital Health **7**, 20552076211033424 (2021). https://doi.org/10.1177/20552076211033426
17. Partners - BD4QoL. https://www.bd4qol.eu/wps/portal/site/big-data-for-quality-of-life/about-bd4qol/partners. Accessed 30 Nov 2022
18. Osoba, D., Rodrigues, G., Myles, J., Zee, B., Pater, J.: Interpreting the significance of changes in health-related quality-of-life scores. J. Clin. Oncol. **16**(1), 139–144 (1998). https://doi.org/10.1200/JCO.1998.16.1.139
19. Cocks, K., et al.: Evidence-based guidelines for interpreting change scores for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. Eur. J. Cancer **48**(11), 1713–1721 (2012). https://doi.org/10.1016/j.ejca.2012.02.059
20. Chan, A.W., et al.: SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann. Intern. Med. **158**(3), 200–207 (2013). https://doi.org/10.7326/0003-4819-158-3-201302050-00583
21. Clinical Trials Transformation Iniciative. https://ctti-clinicaltrials.org/. Accessed 18 Oct 2022
22. Kapoor, A., Nambisan, P., Baker, E.: Mobile applications for breast cancer survivorship and self-management: a systematic review. Health Inform. J. **26**(4), 2892–2905 (2020). https://doi.org/10.1177/1460458220950853
23. Charbonneau, D.H., et al.: Smartphone apps for cancer: a content analysis of the digital health marketplace. Digital Health **6**, 2055207620905413 (2020). https://doi.org/10.1177/2055207620905413

Pervasive Health for COVID-19

COVID-19 Classification Algorithm Based on Privacy Preserving Federated Learning

Changqing Ji[1,2], Cheng Baoluo[2], Gao Zhiyong[2], Qin Jing[3], and Wang Zumin[2](✉)

[1] College of Physical Science and Technology, Dalian University, Dalian 116622, Liaoning, China
[2] College of Information Engineering, Dalian University, Dalian 116622, Liaoning, China
wangzumin@163.com
[3] School of Software Engineering, Dalian University, Dalian 116622, Liaoning, China

Abstract. In order to solve the problems of complex feature extraction, slow convergence of model and most of the deep learning based COVID-19 classification algorithms ignore the problem of "island" of medical data and security. We innovatively propose a COVID-19 X-ray images classification algorithm based on federated learning framework, which integrates hybrid attention mechanism and residual network. The algorithm uses hybrid attention mechanism to highlight high-resolution features with large channel and spatial information. The average training time is introduced to avoid the long-term non-convergence of the local model and accelerate the convergence of the global model. For the first time, we used the federated learning framework to conduct distributed training on COVID-19 detection, effectively addressing the data "islands" and data security issues in healthcare institutions. Experimental results show that the Accuracy, Precision, Sensitivity and Specific of the proposed algorithm for COVID-19 classification on datasets named' COVID-19 Chest X-ray Database' can reach 0.939, 0.921, 0.928 and 0.947, respectively. The convergence time of the global model is shortened by about 30 min. That improves the performance and training speed of the COVID-19 X-ray image classification model with privacy security.

Keywords: Federated learning · COVID-19 · Convolutional block attention module · Resnet50 · Privacy preserving

1 Introduction

Corona Virus Disease-2019 (COVID-19), has been rampant around the world since the outbreak in 2019, and according to WHO statistics, the number of people diagnosed worldwide has exceeded 625 million by October 26, 2022, and the number of deaths has exceeded 6.56 million. A large number of them died due to severe lung infection found late, so it is particularly important to detect COVID-19 patients as early as possible. X-ray has become a key screening method for the early detection of COVID-19 due to its wide penetration rate, low cost and high accuracy [1]. Doctors will be able to analyze the X-ray images to determine if the patient has been diagnosed with COVID-19.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 161–173, 2023.
https://doi.org/10.1007/978-3-031-34586-9_13

Currently, the success of Deep Learning (DL) in image detection has made it an important technology for computer-aided medical applications, which can help doctors analyze chest X-rays to make quick judgments. Deep learning needs a large amount of training data with strong differences to train a deep learning model with high accuracy and strong robustness. However, because the privacy and security of data has been paid more and more attention around the world, various countries have introduced various laws and regulations to regulate and protect data security. In May 2018, the EU issued the General Data Protection Regulation (GPR) Act for data protection; On September 1, 2021, the Data Security Law of the People's Republic of China was officially put into effect. So specific medical data from the hospital is not allowed to be leaked, and collecting training data that meets the requirements is a major challenge.

Federated Learning (FL) [2] is an effective method to solve the problem of data collection. In 2017, it was proposed by H.Brendan McMahan and others from Google. As a privacy preserving distributed machine learning technology, it can deploy deep learning algorithm in each client and use local data to train the model. Then the model is aggregated by the central server and redistributed to each client for training. In this way, it can ensure that the data is not out of local clients, the cooperation between the client training model, and comply with the data security protection related treaties.

Some scholars have noted data security issues. Psychoula I. et al. [3, 4] consider privacy preservation in deep learning. Yao X. et al. [5] found privacy issues of physical objects in the IoT. Some scholars proposed to introduce federated learning for neural network model training. Pfohl et al. [6] proposed distributed privacy learning for EMR in the federated environment, and they further proved that the performance of this model can be comparable to that of centralized training. Dayan et al. [7] used a federal learning framework to provide data of COVID-19 patients from more than 20 institutions around the world. The model input patients' vital signs, laboratory data and chest X-ray. Predict the vital signs of a symptomatic COVID-19 patient from presentation to the next 72 h. Feki et al. [8] proposed a federated learning framework based on deep convolutional neural network (VGG16 and ResNet50) for COVID-19 detection in chest X-ray images. Zhang et al. [9] designed a federated learning system model based on dynamic fusion for the analysis and detection of CT scan or X-ray images of COVID-19 patients. Liu et al. [10] compared the performance of four different network models (MobileNet, ResNet18, MoblieNet and covid) using the federated learning framework.

Aiming at the existing problems and development status of existing models, we propose a COVID-19 X-ray image classification algorithm (FL-Resnet-CBAM) based on federated learning with hybrid attention mechanism and residual network. Mainly has the following three innovations:

1. Proposed distributed training of COVID-19 classification model in federated learning framework to effectively solve the data "island" problem and user privacy security problem in medical institutions.
2. Improve the Resnet network and add the mixed attention mechanism into the Resnet network, which significantly improves the model detection accuracy.
3. Optimize the federated learning framework and introduce the average training time of local models in the training process, which can prevent the global model from not converging for a long time. The AdaGrad optimizer is used to make use of sparse

gradient information in discrete clients to achieve more efficient convergence of the model.

2 Related Work

2.1 Residual Network

He et al. [11] put forward such as not changing input by convolution layer at the same time increase the same mapping makes input directly mapped to the output of the nonlinear layer, namely the residual network. There are five stages in the Resnet50 network, as shown in Fig. 1. Stage1 is mainly a preprocessing operation. Residual modules in Stage2-Stage5 include mapping module Identity Block (Id Block) and Convolution Block (Conv Block). The dimension of input and output vectors of Id Block are the same, and the network can be deepened directly through series. Learn deep semantic information. The dimensions of the input and output vectors of Conv Block are different, and 1 × 1 convolution should be performed to match the dimensions.

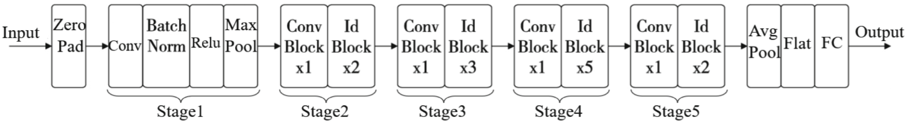


Fig. 1. Resnet50 Network structure

2.2 Hybrid Attention Mechanism

Hybrid attention mechanism refers to the comprehensive evaluation of Channel Attention Module (CAM) and spatial Attention Module (Spartial Attention Module (SAM) simultaneously. Convolutional Block Attention Module (CBAM) [12] is shown in Fig. 2.

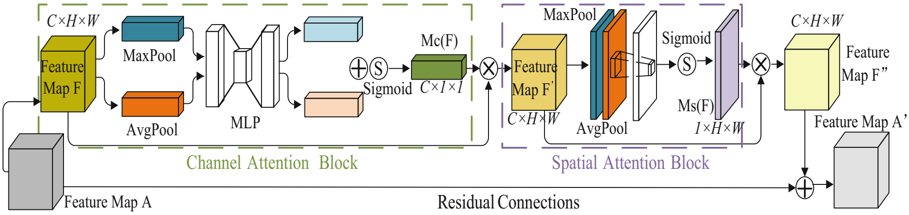


Fig. 2. Hybrid Attention Block CBAM structure

Channel attention module will input feature map F ($H \times W \times C$) through global max pooling and global average poolingbased on width and height respectively to obtain different spatial semantic description operators. Then they are fed into a two-layer shared neural network (MLP) with Relu function as activation function. Then, the feature map output by MLP is added and activated by sigmoid to generate channel attention feature, namely **Mc**(**F**), formula (1). Then, **Mc**(**F**) is multiplied with the input feature map F to obtain the feature map F′ containing channel attention.

$$\mathbf{M_c}(\mathbf{F}) = \sigma\left(\mathbf{W}_1\left(\mathbf{W}_0\left(\mathbf{F}^{\mathbf{c}}_{\mathbf{avg}}\right)\right) + \mathbf{W}_1\left(\mathbf{W}_0\left(\mathbf{F}^{\mathbf{c}}_{\mathbf{max}}\right)\right)\right) \tag{1}$$

σ is sigmoid, $W_0 = \frac{C}{r} \times C,\ W_1 = C \times \frac{C}{r}$.

The spatial attention module takes the feature map F′ output from the channel attention module as the input feature map. Two $H \times W \times 1$ feature maps are obtained by global max pooling and global average pooling operations along the channel dimension. Concat operation is performed on these two feature maps based on channels, and then a convolution operation with a convolution kernel of 7×7 is performed. The dimension is reduced to one channel $H \times W \times 1$, and then the spatial attention feature, namely **Ms**(**F**), formula (2). The input feature map F′ of **Ms**(**F**) module is multiplied to obtain the feature map F″ including channel attention and spatial attention.

$$\mathbf{M_s}(\mathbf{F}) = \sigma\left(f^{7\times 7}\left(\left[\mathbf{F}^{\mathbf{s}}_{\mathbf{avg}}; \mathbf{F}^{\mathbf{s}}_{\mathbf{max}}\right]\right)\right) \tag{2}$$

σ is sigmoid function, and $f^{\ 7\times 7}$ is convolution operation with filter size 7 × 7.

3 FL-Resnet-CBAM Classification Algorithm

3.1 General Framework of FL-Resnet-CBAM

In this paper, an image classification algorithm based on federated learning and deep residual network fusion hybrid attention mechanism (namely FL-Resnet-CBAM) is proposed to realize the recognition of COVID-19 images. The overall algorithm flowchart is shown in Fig. 3.

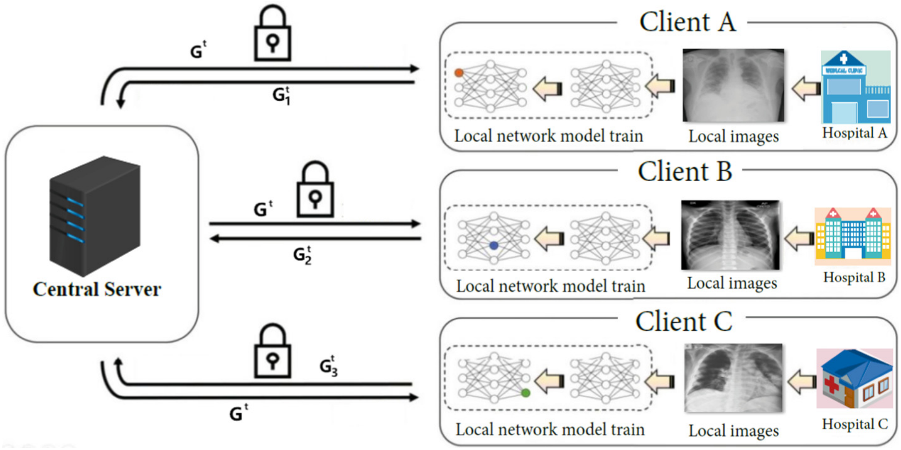


Fig. 3. FL-Resnet-CBAM algorithm framework

In this paper, the centralized aggregation method is adopted to do horizontal federation learning and training, which is mainly composed of aggregation server and multiple discrete clients. Two-way communication can be conducted between server and client, but each client cannot initiate communication. The federated server can use a cloud server, and the local client can be a discrete healthcare facility. The server plays the role of distributing and aggregating models, and the client plays the role of receiving models and training local models. The improved Resnet network model is deployed in each client, and the data used for model training is independent and distributed. The local data in each institution consists of different numbers and types of COVID-19 X-ray images.

3.2 Improved Resnet50 Network

The original Resnet50 residual network, after performing a 5-stage residual convolution operation, was directly and fully connected for classification, but the accuracy on classification of new coronary pneumonia images was obviously not sufficient.

The CBAM module adopts the method of serial channel attention and spatial attention, that is, the feature map F is first corrected by channel attention to F' , and then F' is corrected by spatial attention. Through this module, the model can learn what information in the image and where features need to be emphasized or suppressed, which can effectively help the information flow in the network and enhance feature extraction. The combination of channel attention and spatial attention can save parameters and computational power and can be easily integrated into the existing network architecture.

On basis of the original Resnet50, the mixed-domain attention module CBAM is added, as shown in Fig. 4. Specifically, the CBAM module is embedded in the last convolutional layer of Stage5, and the residual structure is introduced into the model with the embedded attention mechanism, which can effectively suppress the problems of gradient explosion and network degradation. The advantage of embedding CBAM module in the last layer is that the pre-training parameters of Resnet50 model can be used to accelerate the convergence of the training model.

3.3 The Training Flow and Optimization of FL-Resnet-CBAM

In the model training, the initial training parameters are first distributed to all clients, and k clients are randomly selected from the current client set N to participate in the training. In each round of training, the server selects only some clients to participate, which can improve the convergence speed of the local model and accelerate the convergence of the global model.

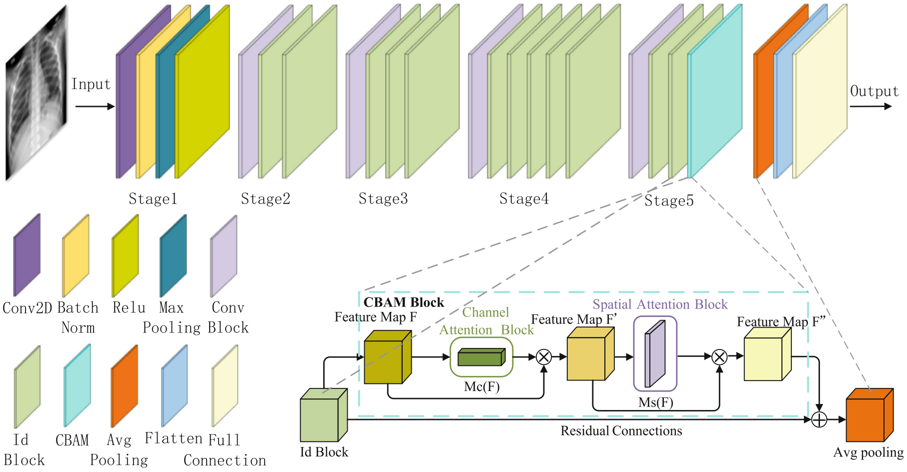


Fig. 4. Improved Resnet50 network structure

Considering the problem that the local model does not converge for a long time due to factors such as uneven data distribution and poor communication of some clients, this paper introduces the average training time of the local model *Tavg,* see (3). In the s round of training, the convergence time of the local client is T^s. In the first round of training, the initial *Tavg* is the fixed value measured in the experiment, and *Tavg* is the average value of the model training time T^s in the first s round.

$$Tavg = \frac{1}{s}\sum_{s=1}^{s} T^s \tag{3}$$

The aggregation server performs one aggregation operation on all model gradients, and we use the FedAvg aggregation algorithm [2].

$$G^{s+1} \leftarrow \sum_{k=1}^{K} \frac{Di}{D} L_i^{s+1} \tag{4}$$

L_i^{s+1} denotes the local model parameters of the *ith* client at round $s + 1$, and G^{s+1} is the global model update parameter (Table 1).

Table 1. FL-Resnet-CABM central server steps.

```
Algorithm 1 FL-Resnet-CABM central server steps
Input: S, K, $G^0$ , $G^s$ , $T^s$ , Tavg
Output: by evaluated the FL- Resnet -CBAM model parameters$G^{s+1}$
Create a task and start training
for global model aggregation count s (1<s< S) do
    Arbitrarily pick k (1 < k < K) clients to federation training
    for the k clients selected in parallel to do
        Call Algorithm 2 to obtain the current training time $T^s$ and the model parameters of each client at the end of the sth round of training$L_i^{s+1}$
        if ($T^s$ > Tavg)
            then Send a non-aggregation command to clients
        else Send aggregation command to clients
        end if
    end for
    Global model parameters $G^{s+1}$ are obtained by FedAvg algorithm, see (4)
    if (evaluate global model convergence = = true)
        then stop the client model training and send $G^{s+1}$ distributed to all clients which aggregation commands are sent
    end if
end for
```

Considering the discrete distribution of each client in the federal learning framework, the traditional gradient descent algorithm is limited in its learning rate adjustment strategy and easily falls into many local suboptimal solutions, so this paper uses AdaGrad (Adaptive Gradient) optimization algorithm [13].

$$g_{t,i} = \nabla_\theta J(\theta_i) \tag{5}$$

$$\theta_{t+1,i} = \theta_{t,i} - \frac{lr}{\sqrt{G_{t,ii} + \in}} \times g_{t,i} \tag{6}$$

$g_{t,i}$ denotes the moment t of θ_i gradient. $G_{t,ii}$ denotes the gradient of the parameter θ_i of the sum of squared gradients (Table 2).

Cross entropy measures the degree of difference between two different probability distributions in the same random variable, and is expressed in machine learning as the difference between the true probability distribution and the predicted probability distribution. The smaller the value of cross-entropy, the better the model prediction.

The Cross Entropy Loss function is often standard with softmax in classification problems. Softmax processes the output so that the predicted values of its multiple classifications sum to 1, and then calculates the loss by cross entropy.

$$\text{Loss} = -[y \times \log \hat{y} + (1 - y) \log\left(1 - \hat{y}\right)\Big] \tag{7}$$

y is the label value 1 or 0, $\hat{y}$ is the probability that the image is predicted to be a new coronary pneumonia image.

Table 2. FL-Resnet-CABM clients steps.

```
Algorithm 2 FL-Resnet-CABM clients steps
Input: D, Di$G^{s+1}$ , i, C, lr, T$^{s}$ , B, momentum,minibatch
Output: local model parameters evaluated by$L_i^{s+1}$ , T$^{s}$
Get the latest model parameters $G^{s+1}$ from the server
for local model training times c(1<c<C) do
    Randomly divide Di into B = Di/minibatch copies
    Get the local model parameters from the previous iteration
    for  b(1<b<B) do
        Get the training parameters state_dict() for each layer from $G^{s+1}$
        Get the lr and momentum to initialize the AdaGrad optimization, see (5,6)
        Call improved Resnet50 model to start training
        Iterate over the data and labels and place them in cuda()
        Optimizer clears the gradient zero_grad()
        Cross entropy cross_entropy calculates the loss value, see equation (7)
        Model gradient update $L_c^{b+1}$ based on parameters in the convolution kernel
    end for
end for
Send this round of client training time T$^{s}$ to central server
if (evaluate local model convergence = = true)
    then get local model parameters updated $L_i^{s+1}$ =$L_C^{B}$ and send it to central server
end if
```

4 Experiments and Analysis

4.1 Experimental Environment

See Table 3.

Table 3. Experimental environment.

Parameters	Configuration
CPU	Intel Core i7-11700 @2.5 GHz
Memory	16 GB
Graphics Cards	GeForce RTX 3080 Ti 12 GB
Operating System	Windows 10
Deep Learning Framework	pytorch 1.5.1
Programming Languages	Python 3.6

4.2 Datasets

We used the dataset COVID-19 Chest X-ray Database [14, 15] to conduct the relevant experiments. We selected 2052 COVID-19 positive cases and 2969 normal lung images considering the model training efficiency and the complexity of data cleaning. The training set consists of 4021 images, 1652 COVID-19 images, and 2369 normal lung images; the test set consists of 1000 images, 400 COVID-19 images, and 600 normal lung images. The image format within this dataset is png with a size of 299 × 299 × 3.

In this paper, we increase the data discrepancy by rotating the images left and right and cropping them at random centers, because the images rotated at large angles will cause some information of the images to be lost, so the images are rotated arbitrarily between 10° and 30° left and right. When the model is trained, an oversampling method is used to have a put-back sampling operation for a small number of image categories, and with this operation method, the impact of data imbalance on the classification model can be reduced, it is necessary to resize the images in the dataset to change the image size to 224 × 224 × 3.

4.3 Experimental Evaluation Metrics

In this paper, four metrics, Accuracy (Acc), Precision (Prec), Sensitivity (Sen), and Specificity (Spec), are used to evaluate the model for image classification (Table 4):

Table 4. Experimental evaluation index.

Evaluation indicators	Prediction Category	Real Category
True Positive (TP)	COV	COV
True Negative (TN)	Nor	Nor
False Positive (FP)	COV	Nor
False Negative (FN)	Nor	COV

$$
\begin{aligned}
Acc &= \frac{TP + TN}{Total} \\
Prec &= \frac{TP}{TP + FP} \\
Sen &= \frac{TP}{TP + FN} \\
Spec &= \frac{TN}{FP + TN}
\end{aligned}
$$

4.4 Experimental Results

4.4.1 Ablation Experiments

Ablation experiments were conducted for the average local model training time proposed in this paper, using the average local model training time within the improved model

FL-ResNet-CBAM denoted as Model A and not using average local model training time denoted as Model B. A total of 10 training sessions were conducted. The training was unfolded with all other training conditions being consistent. The number of global model training sessions and the corresponding convergence times are shown in Fig. 5.

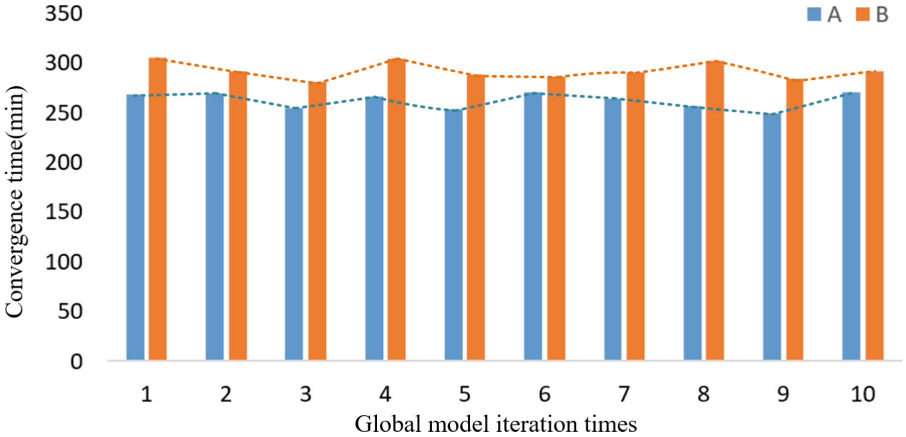


Fig. 5. Convergence time corresponding to the number of training times for Model A and Model B

As can be seen in Fig. 5, the convergence time trend line of model A is lower than that of model B. The average convergence time of model A in 10 global training sessions is 262.3 min, and the average convergence time of model B in 10 global training sessions is 292.5 min, which takes 30.3 min more than model A. It can be seen from the results that the average training time of the local model significantly reduces the convergence time of the global model, which proves the effectiveness of the proposed method.

4.4.2 Comparison Experiments

We designed relevant experiments to test the performance of the proposed model according to the following experimental parameters: N = 10, k = 3, C = 3, S = 20, minibatch = 16, and lr = 0.001. During training, the local training set of the client is randomly divided,it can satisfy the principle of non-independent identical distribution.

In this paper, the VGG16, ResNet18, and ResNet50 models from the VGG [16] and ResNet [11] networks, as well as the ResNet50-SE, a model incorporating the attention module of the SE channel based on ResNet50, and the FL-ResNet50, an improved model of ResNet50 under the federal learning framework, were selected to compare with the proposed method in this paper in the same The prediction results of these six models on the test set are shown in Table 5.

As can be seen from the table, comparing VGG16, ResNet18, ResNet50 can be seen that the model in deepening the depth of at the same time, the model accuracy, precision, sensitivity, specificity are improved, in the residual network, although the number of convolutional layers is increased, but due to the existence of residual units, so that part

Table 5. Comparison of the prediction results of FL-ResNet-CBAM and other models.

Models	Acc	Prec	Sen	Spec
VGG16	0.857	0.850	0.78	0.908
ResNet18	0.871	0.861	0.808	0.913
ResNet50	0.883	0.864	0.840	0.912
FL-ResNet50	0.879	0.868	0.823	0.917
ResNet50-SE	0.914	0.901	0.883	0.935
FL-ResNet-CBAM	**0.939**	**0.921**	**0.928**	**0.947**

of the convolution that does not provide positive feedback is not counted in the results, so The relevant performance results of ResNet50 are higher than those of ResNet18, but not too far apart, and after adding the SE module on the basis of ResNet50, the learning of the model in terms of channels is enhanced, which makes the feature acquisition more comprehensive, and thus the performance of the ResNet50-SE model achieves a comprehensive surpassing of the traditional VGG and ResNet models.

By applying the ResNet50 model to the federal learning framework to achieve the purpose of privacy protection, but from Table 5, we can see that the accuracy and specificity of the FL-ResNet50 model have a small improvement compared with the ResNet50 model, while the accuracy and sensitivity of have decreased, which indicates that the application of deep learning models to the federal learning framework does not necessarily lead to an improvement in all aspects of the model. This suggests that the application of deep learning models in the federal learning framework does not necessarily improve the performance of all aspects of the model, and even degrades some of the performance. This paper argues that this is part of the limitations of federal learning, which may sacrifice a small amount of accuracy at the expense of a significant increase in security.

The FL-ResNet-CBAM model incorporating the CBAM module proposed in this paper achieves both improved prediction results related to the model while using the federal learning framework to ensure security, and all metrics of the FL-ResNet-CBAM model are higher than those of other models, comparing with the FL-ResNet50 model that also uses the federal learning framework, its accuracy and sensitivity are significantly improved compared with the The accuracy and sensitivity of the FL-ResNet50 model are 6 and 10.5 percentage points higher than those of the FL-ResNet50 model, respectively, showing that the CBAM module, which focuses on both channel and spatial information, can achieve the accuracy, precision, sensitivity, and specificity of new coronary pneumonia image classification. This shows that the performance of the model does not degrade under the federal learning framework, but the learning of only the channel information and the lack of spatial information will lose some of the model performance.

During the 20 rounds of global model training, the relationship between the accuracy of each model with increasing number of global iterations is shown in Fig. 6. To show

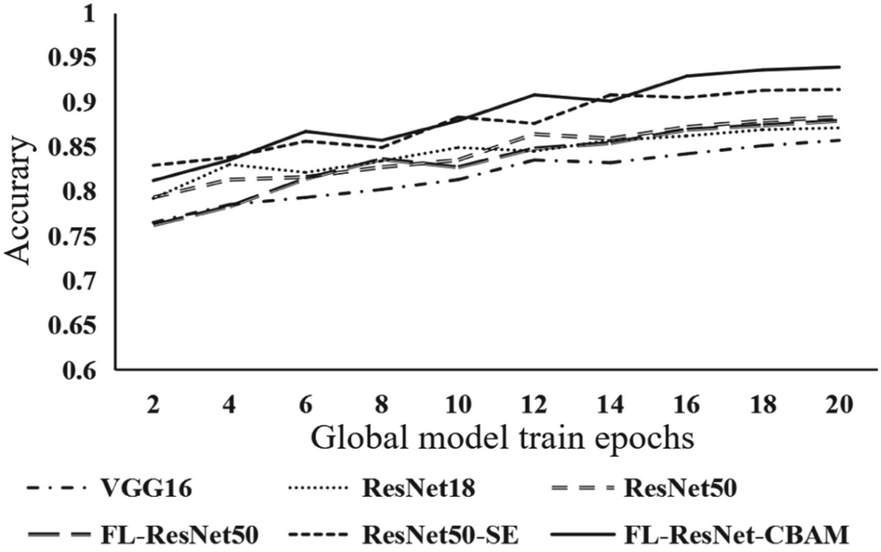


Fig. 6. Relationship between global model train epochs and accuracy of each model

more clearly the difference of accuracy curves of each model in the longitudinal direction, the accuracy starting point of the vertical axis of Fig. 6 is placed at 0.6.

The training process of FL-ResNet-CBAM model is relatively smooth and converges faster than other models, which is partly due to the fact that the federal learning model itself is a distributed model, the new model distributed under each round of client aggregation is normalized, and the local model average training time is used, which allows the model to converge faster.

5 Discussion

In this paper, by improving the residual network ResNet50 under the federal learning framework and adding the hybrid domain attention mechanism, we achieve the improvement of the accuracy, precision, sensitivity, and specificity of the classification of X-ray images of neocoronary pneumonia, and introduce the average training time of the local model in the server-client model training, which solves the problem that the local model does not converge for a long time causing the global model to be unable to converge. The use of AdaGrad optimizer to achieve more efficient model convergence under discrete clients using the information of sparse gradients, and the use of federated learning for distributed training ensures the privacy of medical data and has the advantage of breaking data silos, fully demonstrate the advantages and innovations of the improved algorithm proposed in this paper for new coronary pneumonia detection. In the next work, we consider applying federal learning to more scenarios of detection to break the data barriers while improving the detection accuracy.

References

1. Wong, H.Y.F., et al.: Frequency and distribution of chest radiographic findings in patients positive for COVID-19. Radiology **296**(2), E72–E78 (2020)
2. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics, pp. 1273–1282. PMLR (2017)
3. Psychoula, I., et al.: A deep learning approach for privacy preservation in assisted living. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 710–715. IEEE (2018)
4. Psychoula, I., Chen, L., Amft, O.: Privacy risk awareness in wearables and the internet of things. IEEE Pervasive Comput. **19**(3), 60–66 (2020)
5. Yao, X., Farha, F., Li, R., Psychoula, I., Chen, L., Ning, H.: Security and privacy issues of physical objects in the IoT: challenges and opportunities. Digit. Commun. Netw. **7**(3), 373–384 (2021)
6. Pfohl, S.R., Dai, A.M., Heller, K.: Federated and differentially private learning for electronic health records. arXiv preprint arXiv:1911.05861 (2019)
7. Dayan, I., et al.: Federated learning for predicting clinical outcomes in patients with COVID-19. Nat. Med. **27**(10), 1735–1743 (2021)
8. Feki, I., Ammar, S., Kessentini, Y., Muhammad, K.: Federated learning for COVID-19 screening from chest X-ray images. Appl. Soft Comput. **106**, 107330 (2021)
9. Zhang, W., et al.: Dynamic-fusion-based federated learning for COVID-19 detection. IEEE Internet Things J. **8**(21), 15884–15891 (2021)
10. Liu, B., Yan, B., Zhou, Y., Yang, Y., Zhang, Y.: Experiments of federated learning for covid-19 chest x-ray images. arXiv preprint arXiv:2007.05592 (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
12. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
13. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**(7), 261 (2011)
14. Rahman, T., et al.: Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. Comput. Biol. Med. **132**, 104319 (2021)
15. Chowdhury, M.E., et al.: Can AI help in screening viral and COVID-19 pneumonia? IEEE Access **8**, 132665–132676 (2020)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

Flattening the Curve Through Reinforcement Learning Driven Test and Trace Policies

Andrei C. Rusu(✉), Katayoun Farrahi, and Mahesan Niranjan

University of Southampton, Southampton, UK
{a.rusu,k.farrahi,mn}@soton.ac.uk

Abstract. An effective way of limiting the diffusion of viruses when vaccines are unavailable or insufficiently potent to eradicate them is through running widespread "test and trace" programmes. Although these have been instrumental during the COVID-19 pandemic, they also lead to significant increases in public spending and societal disruptions caused by the numerous isolation requirements. What is more, after the health measures were relaxed across the world, these programmes were unable to prevent substantial upsurges in infections. Here we propose an alternative approach to conducting pathogen testing and contact tracing that is adaptable to the budgeting requirements and risk tolerances of regional policy makers, while still breaking the high risk transmission chains. To that end, we propose several agents that rank individuals based on the role they possess in their interaction network and the epidemic state over which this diffuses, showing that testing or isolating just the top ranked can achieve adequate levels of containment without incurring the costs associated with standard strategies. Additionally, we extensively compare all the policies we derive, and show that a reinforcement learning actor based on graph neural networks outcompetes the more competitive heuristics by up to 15% in the containment rate, while far surpassing the standard random samplers by margins of 50% or more. Finally, we clearly demonstrate the versatility of the learned policies by appraising the decisions taken by the deep learning agent in different contexts using a diverse set of prediction explanation and state visualization techniques.

Keywords: epidemic-control · target-test-and-trace · reinforcement-learning

1 Introduction

The recent pandemic caused by the SARS-CoV-2 virus has fundamentally shaped the way we plan for and respond to the spread of highly-infectious pathogens. Drastic control measures like imposing general lockdowns proved to be particularly damaging to the global economy and the wellbeing of the population [42], causing widespread discontent among all social strata.[1] As such, less restrictive

---

[1] COVID-19 Attitudes Survey by YouGov: https://tinyurl.com/yougov-attitudes.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 174–206, 2023.
https://doi.org/10.1007/978-3-031-34586-9_14

health interventions were introduced in lieu to curb dangerous infection rates, such as educating the public to socially distance, deploying large-scale testing schemes and quarantining contacts through different tracing mechanisms [21]. Despite the advent of highly effective vaccines [3,12], financing and support for these measures continued for several months in the majority of the Western world, fueled by evidence of their continued efficacy [63,83]. However, with the emergence of seemingly milder variants [91], concerns about the limitations [67] or the societal impact [15,60] of the aforementioned interventions, and growing evidence of reduced public compliance [18], several administrations decided to significantly reduce the resources allocated for these programmes. In the United Kingdom, for example, the new "living with COVID" strategy meant appreciable cost reductions could be achieved [61], while heavy disruptions like the recent "pingdemic" could be entirely avoided [76,81]. Unfortunately, blindly scaling down the public health efforts to break transmission chains has proven unsuccessful as cases across the country soared yet again within a relatively short timeframe,[2] trend that has been replicated across Europe [34]. With the vaccine protection waning over time [24,55], and with demand for further doses decreasing among healthy adults [94,97], similar surges could reoccur henceforth.

In this work, we propose a major shift in the implementation of "test and trace" programmes that is adaptable to a country's budget and risk tolerance, while minimizing the burden of viral infection chains. To achieve this, we study different types of targeted policies for conducting testing and isolating contacts in an epidemic under fixed budgeting requirements, and show that a reinforcement learning agent can derive powerful and generalizable policies that outperform all baselines considered in terms of infection reach. We validate our results on several epidemic, budget and interaction network configurations, illustrating the versatility of our proposed method. Moreover, we demonstrate that even static non-learning agents significantly outcompete customary untargeted strategies.

The contributions in this paper are threefold:

1. We put forward a novel way of operating public health interventions in a realistic scenario where economic and societal disruptions are to be minimized: restricting the testing and tracing efforts to higher-risk individuals. To that end, we derive highly-effective policies using different agents, including centrality-based, neighborhood-based and learning-based, comparing them against more traditional approaches, such as random, acquaintance or frequency-based sampling. Our reinforcement learning agent, backed by a Graph Neural Network (GNN) adapted from the recent development of [65], is shown to outperform the other methods in both tasks across numerous configurations, despite being trained using a simple test prioritization setup with partial observable information.
2. Aside from presenting the numerical results and epidemic curves resulted from running our control policies over multiple simulations, we also study what the learning-based agent chooses to focus on while making its decisions. For such a system to be deployed in the real world, policy makers

---

[2] COVID-19 Infection Survey by ONS: https://tinyurl.com/ons-covid19.

need to be reasonably confident the model produces sensible outputs. At the same time, testing or isolation decisions have to be explainable and verifiable when audited or contested. Here, we explore a perturbation-based technique for explaining the policy derived by an agent's GNN module, GraphLIME [39], putting into perspective the former's superior adaptability. Moreover, we propose visualizations for the inferred node embeddings that can be used to direct community-wide interventions or scrutinize the model's performance.

3. We apply our framework to several scenarios featuring COVID-specific spreading models, including a multi-site mean-field [40,83] and an agent-based model, with parameters obtained from [20], and show that our agents consistently perform well across a diverse set of experimental setups.

2 Related Work

2.1 Epidemic Modelling

Traditionally, simulating epidemics has been accomplished using either equation-based or agent-based models. The first of these is possibly the most common, owing its appreciable success to early work by [45], where the modelled population was said to transition between disease-specific compartments according to a system of ordinary differential equations. Recent years, however, have seen agent-based approaches become more popular, partly due to their superior granularity and ability to assess a system's behavior at the individual level [98]. Government-advising groups in the United Kingdom employed this paradigm during the initial waves of the COVID-19 pandemic to assess the effects of public health interventions [25,35]. Others used such formulations to study the combined effects of manual tracing with digital solutions at various application uptakes, employing parameters fitted to infection data from several regions [1,83]. In this study, we simulate viral epidemics using a modified version of a recently-proposed multi-site mean-field model [83], which relies on the SEIR compartmental formulation but retains the capacity to leverage an individual's locality information through contact graphs and mean effects [23,40]. For completeness, we also investigate our policies in a purely agent-based setup, similar in spirit to the network-based approaches proposed in recent works [1,65]. In both cases, we employ the COVID-specific dynamics parameters inferred by [20], and allow all disease-unrelated events to be time-discretized (i.e. selecting an action or updating the active links set takes place every t_u days, with t_u=1).

2.2 Graph Neural Networks and Reinforcement Learning

A few years back, graph neural networks became one of the de facto machine learning tools for processing graph-structured information [27,110]. The earliest studies in this space defined a GNN as a set of two functions: transition f_θ and output o_θ [30,86]. The former expresses the dependence between a node i and its vicinity, while the latter controls the space spanned by the model output.

These functions form the basis of what later came to be known as the message-passing paradigm [29], which quickly became dominant in the field due to its effectiveness and computational efficiency [11]. After graph convolutional networks (GCN) were first introduced [48], many GNN successes followed course [11,32]. Motivated by the accomplishments of attention mechanisms in natural language processing [19,104] and computer vision [4,46], authors soon enhanced GNNs with attention capabilities, often increasing their performance (e.g. GAT [105], GATv2 [10]). Expressive power bounds for the message passing algorithm were later noted by [108], who then proposed an architecture that reaches the upper limit of the widely-used Weisfeiler-Lehman heuristic (1-WL): the Graph Isomorphism Network (GIN). Efforts to break node symmetries and surpass this upper bound have been significant ever since, with many approaches currently existing: augmenting the nodes with random features [85], modifying the message passing rule [6], or changing the input graph structure itself [71]. Additionally, issues such as feature *oversmoothing* [36,74] and *bottlenecks* [2] have been identified as common reasons for underperforming message passing systems, with proposed solutions ranging from maintaining a low layer count and connecting all nodes in the last layer to ease information flow, to augmenting the message exchange routine (e.g. Neural Sheaf Diffusion [8]). Our framework leverages ideas from GATv2 and GIN to attain expressive power and computational efficiency, while reducing the impact of the above problems by using randomised node features and a small number of GNN layers, with a final fully-adjacent layer to mitigate the over-squashing of long-range dependencies.

A widely-used approach for explaining predictions in deep learning involves perturbing the inputs and fitting local explainable models to each data point and its corresponding perturbations. LIME [80] and SHAP [58] are two popular examples of this methodology. Although the above are directly applicable to GNNs, they do not possess the capability to leverage structural information from graph data or capture nonlinear relationships between the inputs and the outputs. To solve these limitations, GraphLIME was proposed [39]. GraphLIME replaces the local perturbations matrix with stacked node features selected from a node's neighborhood, fitting nonlinear interpretable models using HSIC Lasso [109].

Among many other domains, GNNs have also been extensively used in the context of epidemiology. From the literature dedicated to COVID-19, we note here several noteworthy efforts: infection forecasting [44,75], full population state estimation [102], finding "patient 0" [90], and controlling public interventions, such as testing [65] or vaccination policies [41].

Sequential decision processes are often modelled via Markov Decision Processes (MDPs) of the form ($\mathcal{S}$, $\mathcal{A}$, $\mathcal{P}$, $\mathcal{R}$) [78], where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, $\mathcal{P}$ is a transition probability matrix, while $\mathcal{R}$ is a reward function for the state-action pairs. Agents sample actions from their policy $a_t \sim \pi(a|s_t;\theta)$, with $a \in \mathcal{A}$ and θ a parametrization, then execute them, transitioning to different states s_{t+1} and earning rewards R_t, according to the environment's $\mathcal{P}$ and $\mathcal{R}$. The goal of reinforcement learning (RL) is to solve MDPs by predicting and maximizing the γ-discounted returns of future rewards $G_t = \sum_{i=1}^{T} \gamma_{t+i}^{i-1} R_{t+i}$ [100].

This is routinely achieved through supervision using a w-parameterized model $V(s_t|w)$ that predicts G_t, and the returns or some intermediary estimates as targets. The first of these approaches is called the Monte Carlo (MC) algorithm, and is known to be effective despite presenting several drawbacks: slow offline learning and high variance [99,100]. For example, a variation of MC featuring search trees was used to derive competitive policies in 2-player board games [92,93]. In contrast, the online temporal difference (TD) learning method casts the sum of the current estimate of the next-step return $G_{t+1} = V(s_{t+1}|w)$ and R_{t+1} as a regression target, lowering the variance and speeding up training [103]. The latter constitutes the basis for many RL algorithms to date, such as the on-policy SARSA [82,100] and the off-policy Q-learning [106], which proved successful in multiple problem instances: reaching or outperforming human-level performance in games [69,70], autonomous car driving [49], and many others.

Approaches that directly optimize both θ and w are called actor-critics [52], and have become the preferred algorithmic choice when faster convergence rates are sought after and sample efficiency is not required. Recent years have seen actor-critic methods like the Proximal Policy Optimization (PPO) [88] and Deep Deterministic Policy Gradient (DDPG) [56] achieve state-of-the-art results across a wide range of challenging tasks [54,88]. Although online implementations are possible, these agents have traditionally been trained using MC.

Learning policies in environments with combinatorial action spaces such as ours has generally been considered a difficult undertaking. In spite of this, RL methods proved to be effective in instances like multiple item [96] or thread popularity selection [33]. In the context of epidemics, an RL system based on multi-armed bandits and demographics data was recently introduced by the Greek authorities to prioritize the COVID-19 testing allocations at border control [5]. For classic combinatorial problems, such as the travel salesman (TSP) and its vehicle routing variants, RL approaches have also been shown to perform well [7,53]. Incorporating graph embeddings into the RL agents have generally lead to improved solvers, outcompeting other learning methods [17,43].

2.3 Influencing Graph Dynamics

The problem of influencing diffusion processes over networks has been studied in many different settings before, most notably for solving influence maximization [73], optimizing immunization strategies [77], and targeting pathogen testing [66]. It has been long established that random vaccination policies tend to be sub-optimal, and even simple heuristics like acquaintance sampling can outperform them [16,68]. Centrality-based strategies were also explored in this context, with PageRank [14], eigenvector [62] or betweenness centrality [84] becoming popular choices. For influence maximization, degree-based strategies were shown to render competitive results (e.g. LIR [57], degree discount [13]). Over time, however, multiple authors have identified problem instances where any centrality measure used by itself can lead to suboptimal results [9,77]. The question of which heuristic to use for what problem has since become a focal point in many application domains. As an alternative, reinforcement learning techniques have been proposed for mixing different heuristics in an optimal manner, thus reducing the impact of the aforementioned drawbacks [65,101]. Node targeting for detecting the state of a spreading process is a slightly less explored use case of control in the literature, but efficient heuristics that exploit the known state of a vertex's neighborhood instead of centrality-derived information have proven to be successful [66]. The domain of prioritizing contact tracing, however, remains largely uninvestigated to date, but recent work suggests that isolating subsets of individuals based on the frequency of appearing in the vicinity of positive cases can lead to similar levels of containment as naively isolating every contact [51].

Meirom et al. introduce a reinforcement learning model that can derive general control policies for diffusion processes over networks, using test prioritization and influence maximization as illustrations [65]. A GNN-based controller, cast in an actor-critic framework, learns effective policies using simulated data, integrating local and long-distance information over time. The elegance of the approach stems from the fact that the training process is not conditioned on having the full epidemic state made available to the agent. The work also shows that it is possible to learn a policy on small networks (e.g. 1000) and deploy it on larger graphs with similar statistics (e.g. 50000, the size of a small city). Our study builds on top of this versatile control framework, but differs from the aforementioned work in several key aspects: First, we extend the problem formulation to cover prioritizing both testing and tracing, amending the framework to accommodate ranking nodes from eligible subsets. The latter also enables us to add a simple extension to all our agents which empirically improves performance: restricting the action space to exclude recently-tested negative individuals. Second, we analyze the control outcomes more thoroughly, looking at longer evaluation episodes than 25 days, plotting epidemic curves, and interpreting the agents' decisions using a perturbation-based explainability technique designed for graphs, GraphLIME. Third, we employ COVID-specific spreading parameters and analyze the behavior of the policies beyond agent-based modelling. Finally, we perform a range of algorithmic changes in our implementation to improve efficiency: using bootstrapping and eligibility traces to mitigate the memory cost of the offline PPO routine, a shared network between the actor and the critic [88] to enrich the graph embeddings, a GATv2 layer in the diffusion module to enable a better tracking of the point-to-point spreading process, multiple GIN layers followed by a final fully-adjacent one in the information module to increase its expressive power, as discussed above, and standard scaling for bounding the exploding node hidden states instead of L^2 normalization or GRU-based transformations.

3 Methodology

3.1 Simulating Epidemics

We simulate several epidemics using the SEIR compartmental model together with COVID-specific parameters obtained from [20]: a base infection rate of $b = 0.0791$ and an average exposed duration of $e = 3.7$ days. In order to remove stochastic artefacts that may conceal performance differences, most of our setups assume that nodes remain infectious for the whole duration of the episode unless they get isolated (i.e. recovery rate $\rho = 0$). Intuitively, the impact of this assumption becomes significant only when the problem becomes *oversaturated* (i.e. the testing budget k and/or the recovery rate ρ are large enough for any agent to achieve containment). For completeness, however, we also present results when ρ is varied (see Section 4.1).

The viral infection diffuses over different interaction network configurations, with events getting generated by either a multi-site mean-field (see [83]) or an agent-based model. These configurations correspond to both common artificial generation methods, such as Erdős–Rényi [22], dual Barabási-Albert [72] or Holme-Kim [38], and real interaction patterns. In practice, the graph connections needed for conducting such fine-grained control would have to be inferred from a monitoring system, like a digital tracing mechanism [65] or human mobility tracking via GPS [89], process which requires careful data anonymization. We assign a transmission weight $w_j \sim \mathcal{U}(0.5, 1)$ to every edge j in our graphs, calculating an interaction's transmission probability by scaling w_j with the base factor b. In the multi-site mean-field simulations, stochasticity is ensured by the events sampling procedure, which is efficiently performed using Gillespie's algorithm [28]. In contrast, the agent-based model relies on sampled exposed-state and recovery durations for each node, $d_i \sim \mathcal{N}(e, 1)$, $r_i \sim \mathcal{N}(\frac{1}{\rho}, 1)$, and the w_j weights to induce variability among individuals. For further details, please consult Appendix A.2.

3.2 Control Setup

Each epidemic is allowed to progress until at least c_a days have passed since the simulation began and a minimum of c_i nodes become infected before the agent commences its interventions. In the first day of control, the agent is informed at random about the status of a proportion c_k of the *infected* population, after which it is only allowed to test k individuals and isolate k_c contacts of recently-detected positive nodes (i.e. in the previous 6 timestamps) per day. As the actor is not aware of a node's state unless it is a part of c_k or it got tested recently, the environment is partially observable. In this work, we fix $c_a = 5$, $c_i = 5\%$ and $c_k = 25\%$, while the budgets are varied between experiments. A block diagram of our framework, which includes the agents' class hierarchy, is provided in Fig A1.

During evaluation, each agent is asked to select the top-k nodes to test and the top-k_c contacts to isolate every day, according to their appraisal of the epidemic and graph states. Consequently, this constitutes an instance of the subset selection problem [79], where nodes that are traced by the system or found to be positive are marked as isolating, becoming incapable of infecting other nodes. In principle, those individuals remain disconnected from the graph, yet we allow messages to continue flowing through their connections during the training phase of the learning-based agents. Importantly, the process of tracing is assumed to be carried with delays shorter than a day, which usually implies that a contact tracing application is already deployed and functioning [26,107]. To evaluate the efficacy of each policy, we analyze the fraction of nodes kept healthy through the entire epidemics and the corresponding infection curves.

3.3 Baseline and Learning Agents

In this study, we consider a wide variety of baseline agents for controlling the viral diffusion that leverage separate heuristics: Random samplers (or *randag*); Acquaintance samplers (or *acq*); Centrality-based (e.g. Degree or *deg*, Eigenvector or *eig*, PageRank or *prank*, Closeness, Betweenness); Neighborhood state-based (or *neigh*). The latter is the only baseline that uses information about the epidemic state, targeting the nodes that have the highest number of positively-detected neighbors in their 2-hops vicinity via lexicographical ordering [66].

Aside from the above, when ranking the contacts of positives, additional information can be exploited through heuristic methods: the frequency with which nodes appear in the neighborhood of detected cases. We derive two baselines from the above: Frequency, which randomly samples nodes with probabilities proportional to the individual frequencies (equivalent to the tracing mechanism studied in the multi-site mean-field approach of [83]), and Backward, which greedily picks the nodes with the highest frequencies (as per [51]).

We also propose a simple yet powerful extension to these baselines: recollection of recent negative test results. This effectively restricts the action space to untested nodes in the past t_n days, speeding up the network exploration. We set $t_n = 3$, an appropriate timeline for COVID-19 [95], which renders good results empirically.

Our learning-based agents are inspired by the recent publication of [65], leveraging multiple GNNs due to their proven efficacy for targeting testing campaigns. The abstract structure of our models remains similar to the previous work, with a single-layered diffusion module and a long-range information module, followed by two multi-layer perceptrons (MLPs), one that computes the node hidden states h_i, and another that defines the output space. However, our proposed solution features several improvements or simplifications: First, we utilize two output MLPs to produce a score for each vertex and a full state score from the same model, thus sharing the embedding space between the two. Second, we employ a GATv2 layer in the diffusion module to leverage attention when aggregating information from the immediate neighborhood of each node, and 3 GIN layers followed by a fully-adjacent layer in the information module to improve the expressivity and long-range information flow. Finally, after experimenting with different normalization schemes to mitigate the issue of the exploding hidden states h_i (problem also outlined in the aforementioned study), we propose the usage of standard scaling, which leads to stable training behaviors.

In addition to the above, we carefully scrutinised different combinations of node features, choosing the following final set for training our policies: the degree and eigenvector centralities, the number of infected vertices in the 1-hop and 2-hop neighborhoods, 5 random features that break structure symmetries, and 4 test-state features: a one-hot vector of size 3, marking the test status of node i at the previous timestamp (untested, negative or positive), and a binary value marking whether the vertex has ever tested positive. To allow for the hidden states to incorporate information from these features before the training commences, we disable gradient updates for the first 11 passes.

The ranking of nodes can be performed by both a supervised learning (SL) and a reinforcement learning (RL) agent, with little to no changes to the underlying neural network architecture. The SL agent is trained as a simple node classifier by optimizing a binary cross-entropy loss on the infection status of each vertex, with the output space representing the next-step infection likelihood. In contrast, our RL agent gets optimized via a surrogate PPO objective, which only needs access to the total number of infected at each time point (for more details, refer to Appendix A.3), ultimately solving for the criterion below, where E(t), I(t) and R(t) are the number of individuals in each compartment at time t:

$$\min \sum_{t=t_0}^{\infty} \gamma^{t-t_0} (E(t) + I(t) + R(t)) \tag{1}$$

Here, two reward functions can be used: negative of the number of infected or the number of susceptible vertices at time t (corresponding models denoted as *rl* and *rlpos*, respectively). The performance between the two varies due to numerical reasons, but the differences are small (see Fig A5). Consequently, Section 4 features only the former in the summary tables.

To ensure sufficient exploration during training, the RL agent passes the raw outputs of the ranking model through a softmax function that features a decaying temperature, starting from $\epsilon = 0.5$. Note that other strategies are also possible here, including the transforms proposed in [64] and [65], but our simple alternative proved sufficiently effective at exploring the state space. During evaluation, the sampling process is turned off, greedy actions are taken instead, and the edges connected to positively-identified vertices are masked before being fed to the information module, limiting feature *oversmoothing*. In contrast, we allow the single-layer diffusion GNN to utilize to the aforementioned links such that the positive-related node features can pass through to their neighbors.

By comparing the training behavior of the SL and RL agents with the containment achieved by the centrality-based actors with recollection, we observe a clear distinction between the two, as reflected by Fig 1. While the RL policy outperforms all baselines in several episodes, despite not entering evaluation mode as of yet (i.e. when exploration would be turned off), the SL policy struggles to compete. Further evidence of the SL agent's underperformance can be seen in the plots of Fig A2, as well as in the extensive comparison previously conducted by [65]. Consequently, we focus our main analysis in Section 4 on the policies derived by the RL actors, comparing them against the rest of the baseline agents.

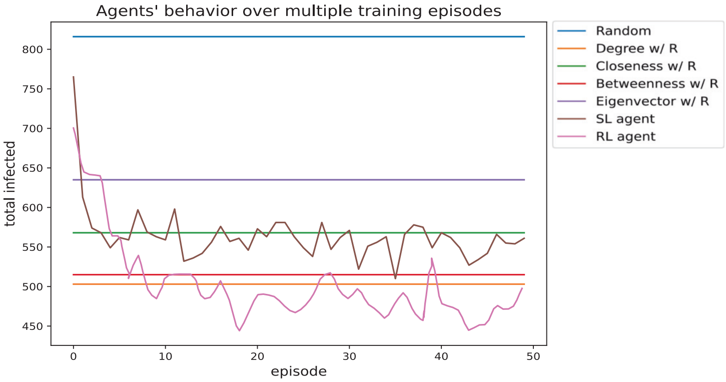


Fig. 1. The learning agents' training behavior. Results obtained by the centrality-based agents and the random tester are plotted for comparison.

4 Results and Discussion

4.1 Prioritizing COVID Testing in Static Graphs

We first explore our agents' policies in the context of targeted testing campaigns. To that end, we investigate the fraction of nodes kept healthy throughout various epidemics triggered across different network models when the budget of daily testing k is fixed, while k_c is set to 0. Once tested positive by the framework, a node gets isolated and eventually acquires immunity, thus remaining uninfectious until the end of the simulation. As stated previously, most of our setups assume nodes do not spontaneously become uninfectious (i.e. $\rho = 0$), but for completeness we present results for different full-recovery rates in Table 1.

Despite being trained for only 50 episodes on a single epidemic configuration spanning a preferential attachment network of 1000 nodes, our reinforcement learning agent consistently outperforms the other baselines across a range of different network sizes (see Table 2), budgets (see Fig A4), and wiring configurations (see Fig A5). Interestingly, as previously hinted by [65], the learning-based agents poses a great generalization capability when the daily budgets scale with

Table 1. Fraction kept healthy with budget $k = 1\%$ and different recovery rates. Average over 5 seeded runs for each of the considered 5 realizations of Barabási-Albert networks with $N = 1000$ nodes and a mean degree of approximately 3. "w/R" denotes agents with recollection of recent negative test results.

Agents	ρ **= 0**	ρ **= 0.01**	ρ **= 0.02**	ρ **= 0.03**
Degree	0.555 ± 0.027	0.616 ± 0.034	0.662 ± 0.039	0.697 ± 0.039
Degree (w/R)	0.744 ± 0.032	0.769 ± 0.028	0.801 ± 0.028	0.847 ± 0.025
PageRank (w/R)	0.720 ± 0.026	0.755 ± 0.023	0.792 ± 0.037	0.834 ± 0.039
RL	**0.822 ± 0.033**	**0.846 ± 0.026**	**0.876 ± 0.026**	**0.897 ± 0.026**

Table 2. Fraction kept healthy with budget $k = 1\%$ and different population sizes. Average over 5 seeded runs for each of the considered 5 realizations of Barabási-Albert networks with a mean degree of approximately 3. "w/R" denotes agents with recollection of recent negative test results. Here, a single model is trained for 50 episodes on a network of size 1000, but its policy is able to generalize to appreciably larger graphs.

Agents	**N = 500**	**N = 1000**	**N = 2000**	**N = 5000**	**N = 20000**
Degree	0.533 ± 0.037	0.555 ± 0.027	0.552 ± 0.017	0.567 ± 0.027	0.557 ± 0.005
Degree (w/R)	0.731 ± 0.031	0.744 ± 0.032	0.736 ± 0.027	0.737 ± 0.025	0.746 ± 0.009
PageRank (w/R)	0.709 ± 0.019	0.720 ± 0.026	0.724 ± 0.021	0.729 ± 0.021	0.725 ± 0.008
RL	**0.817 ± 0.032**	**0.822 ± 0.033**	**0.811 ± 0.024**	**0.821 ± 0.025**	**0.803 ± 0.026**

Table 3. Fraction kept healthy for 1000 nodes. Results are averaged over 5 runs for each of the 5 realizations of a configuration model built using real tracing statistics.

Agents	k **= 20**	k **= 50**
Acquaintance (w/R)	0.465 ± 0.086	0.736 ± 0.085
Degree (w/R)	0.406 ± 0.020	0.746 ± 0.025
Eigenvector (w/R)	0.186 ± 0.013	0.409 ± 0.026
PageRank (w/R)	0.363 ± 0.016	0.668 ± 0.039
RL	**0.506 ± 0.029**	**0.831 ± 0.047**

the number of nodes, making possible a deployment into larger networks, irrespective of the training graph size and artificially without losing efficacy.

Several epidemic curves corresponding to prioritizing testing in 5000 nodes graphs are shown in Fig A3. We note the random approaches perform strikingly poorer than all our informed policies, while the impact of recollection is apparent. Moreover, in spite of using recollection, the heuristics considered remained inferior to the RL policy in terms of the average containment rate.

4.2 Prioritizing Testing in Dynamic Graphs

In the previous section, we analyzed scenarios in which the connections between nodes remain fixed for the entire simulation. However, in practice, the interaction patterns change over time. In Fig 2, we present boxplots of the percentage of nodes kept healthy obtained by different agents on several preferential attachment networks whose active edges are sampled every day (a uniform random fraction is sampled daily from $\mathcal{U}[0.4, 0.8]$). The reinforcement learning agent was retrained to accommodate this dynamic context, allowing the model to pass messages through the most recent edges only. The top performing policies were also evaluated on dynamic networks built using statistics from a real contact tracing network [65], the resulting average containments being displayed in Table 3.

4.3 Targeted Test and Trace Programmes

Next, we investigate the extent to which different combinations of agents tasked with conducting testing and contact tracing under the constraints of a fixed

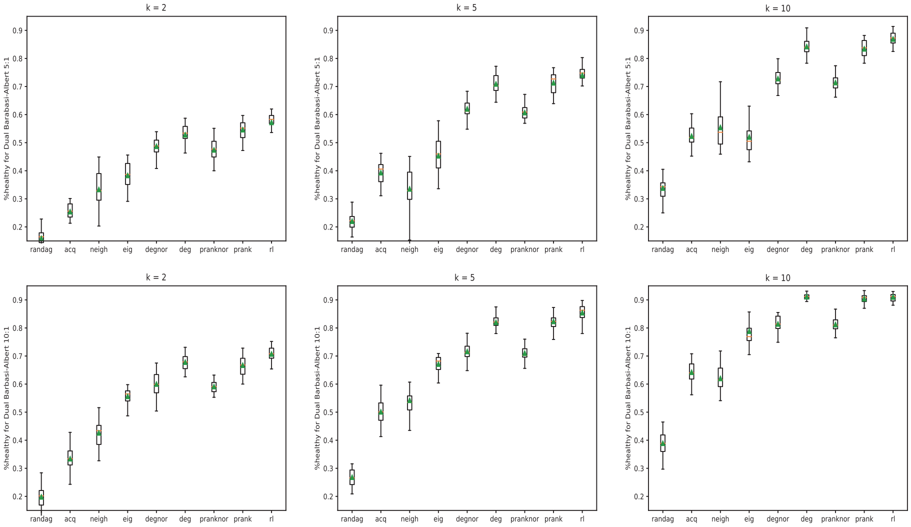


Fig. 2. Infection control performance on different dynamic network architectures. The uncertainties are shown as boxplots.

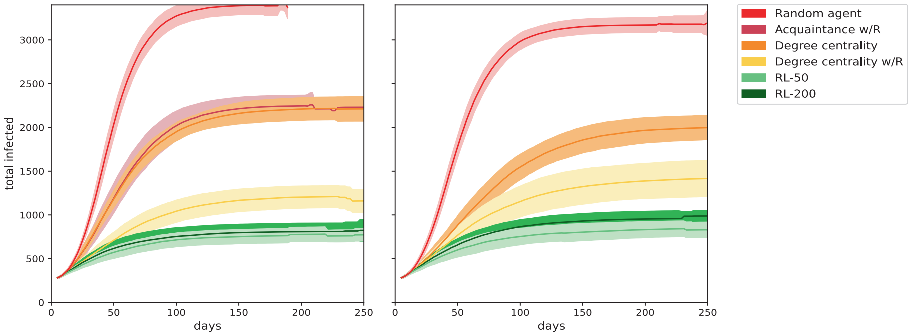


Fig. 3. Averaged epidemic curves and their standard deviations during test and trace control. These are for 5000 nodes Barabási-Albert networks featuring a mean degree of approximately 3, with a daily testing budget of $k = 1\%$ and no tracing on the left, and $k = 10$ with a limit of $k_c = 25$ traced contacts on the right. Two RL agents are displayed: one trained for 50, and the another for 200 episodes.

budget can reduce the spread of a pathogen. For this problem, we train an RL agent for 200 episodes on the same testing task as before, and compare the resulting policy against the other baselines. Tables 4 and 5 confirm the RL tester improves the overall quality of the test and trace programmes, irrespective of the chosen tracer. That being said, employing the same agent to perform the ranking of contacts as well generally improves the containment.

We also inspect the averaged epidemic curves associated with these targeted test and trace campaigns when $N = 5000$. The results obtained by each agent

Table 4. Percentage of nodes kept healthy for graphs of size 1000 and an approximate mean degree of 3, with budgets $k = 2$, $k_c = 5$. Averages over 5 runs for each of the considered 5 realizations of the following: dual Barabási-Albert with $m1 = 5$, $m2 = 1$ (BA 5-1) and $m1 = 10$, $m2 = 1$ (BA 10-1), Holme-Kim (PC), and Erdős–Rényi (ER).

Agents (Test + Trace)	**BA 5-1**	**BA 10-1**	**PC**	**ER**
Random + Random	0.389 + 0.044	0.446 ± 0.060	0.131 ± 0.023	0.199 ± 0.023
Random + Frequency	0.387 ± 0.033	0.465 ± 0.059	0.202 ± 0.030	0.195 ± 0.022
Acquaintance (w/R) + Random	0.541 ± 0.054	0.657 ± 0.054	0.212 ± 0.031	0.215 ± 0.017
Acquaintance (w/R) + Frequency	0.582 ± 0.055	0.674 ± 0.059	0.228 ± 0.040	0.217 ± 0.021
Acquaintance (w/R) + Backward	0.591 ± 0.056	0.769 ± 0.080	0.213 ± 0.039	0.208 ± 0.019
Acquaintance (w/R) + RL	0.644 ± 0.048	0.806 ± 0.058	0.248 ± 0.038	0.217 ± 0.018
Degree (w/R) + Degree	0.764 ± 0.038	0.915 ± 0.032	0.528 ± 0.053	0.333 ± 0.037
RL + Random	0.818 ± 0.034	0.882 ± 0.026	0.542 ± 0.050	0.438 ± 0.043
RL + Frequency	0.832 ± 0.035	0.890 ± 0.033	0.567 ± 0.054	0.448 ± 0.048
RL + Backward	0.849 ± 0.033	0.923 ± 0.023	0.590 ± 0.058	0.434 ± 0.047
RL + Degree	0.853 ± 0.034	0.928 ± 0.014	0.614 ± 0.055	**0.453 ± 0.039**
RL + RL	**0.876 ± 0.025**	**0.936 ± 0.009**	**0.620 ± 0.050**	0.451 ± 0.039

Table 5. Percentage of nodes kept healthy when controlling epidemics over a dynamic real interaction network of 74 vertices, derived from the Social Evolution dataset [59]. Averages over 5 runs for each of the considered 5 infection seeds. Test budget is $k = 2$.

Agents (Test + Trace)	$k_c = 2$	$k_c = 4$
Random + Frequency	0.511 ± 0.130	0.659 ± 0.114
Acquaintance (w/R) + Frequency	0.494 ± 0.113	0.649 ± 0.089
Acquaintance (w/R) + Backward	0.522 ± 0.115	0.654 ± 0.126
Neighborhood (w/R)	0.620 ± 0.108	0.704 ± 0.107
Degree	0.614 ± 0.107	0.741 ± 0.084
Degree (w/R)	0.636 ± 0.104	0.750 ± 0.084
RL	**0.711 ± 0.089**	**0.773 ± 0.069**

is shown on the second column of Fig 3, with the first serving as a test-only reference (i.e. values from Fig A3). As stated before, heuristics with recollection bring large improvements over random policies, yet the RL agents outcompete them in most setups. Note the performance of $k = 50$ tests is similar to $k = 10$ tests, but tracing up to $k_c = 25$ contacts daily. While the balance between these will depend on various factors, the results highlight the effectiveness of tracing.

4.4 Agents Interacting with Different Spreading Dynamics

To assess the ability of the agents to generalize to other spreading dynamics, we compare their achieved containment rates recorded with both a multi-site mean-

Table 6. Fraction kept healthy for 2000 nodes and an average degree of 3. Results represent averages over 5 runs for each of the considered 5 instances of a dual Barabási-Albert model ($m_1 = 10$, $m_2 = 1$). Testing budget is $k = 2$ and no tracing is conducted.

Agents	**Multi-site**	**Agent-based**
Random	0.164 ± 0.037	0.195 ± 0.034
Acquaintance (w/R)	0.251 ± 0.033	0.263 ± 0.035
Degree	0.390 ± 0.032	0.394 ± 0.029
Degree (w/R)	0.443 ± 0.032	0.457 ± 0.034
RL	**0.468 ± 0.035**	**0.477 ± 0.034**

Table 7. Fraction kept healthy for 2000 nodes and an average degree of 3. Results represent averages over 5 runs for each of the considered 5 instances of a dual Barabási-Albert model ($m_1 = 10$, $m_2 = 1$). Budgets are $k = 2$ and $k_c = 10$.

Agents (Test + Trace)	**Multi-site**	**Agent-based**
Random + Random	0.372 ± 0.035	0.371 ± 0.042
Acquaintance (w/R) + Backward	0.633 ± 0.046	0.627 ± 0.053
Degree (w/R) + Degree	0.841 ± 0.034	0.809 ± 0.028
RL + Backward	0.867 ± 0.029	0.851 ± 0.030
RL + Degree	0.889 ± 0.026	0.856 ± 0.025
RL + RL	**0.911 ± 0.020**	**0.882 ± 0.018**

field and an agent-based model run with similar hyperparameters. The RL agent retains all the learned parameters inferred from the previous experiments.

Despite the fact that the control mechanism in the mean-field case relies on discretizing a continuous-time process, we observe minor differences between the two simulation approaches (Tables 6 and 7). This confirms that the agents continue to perform well irrespective of the underlying dynamics.

4.5 Explaining and Visually-Inspecting the Learning Agent's Policy

To derive explanations for the decision taken by our reinforcement learning policy, we employ the GraphLIME algorithm, fitting multiple interpretable models to the raw action-values the model outputs. Fig 4 presents the feature importances derived by GraphLIME for a given day in the early stages of an epidemic, highlighting that the RL agent preferentially attends to the centrality features when it does not possess enough information about the diffusion state. As soon as the tester records positive individuals in the vicinity of a vertex, the rank of the latter increases. After neighborhoods become filled with known infections, the agent targets the affected sectors by focusing on the epidemic state features (see Fig 5). As previous results also suggest, the degree remains an effective predictor for a node's importance throughout the process. Interestingly, the *untested* flag is often correlated with the action scores, which may indicate the agent favors exploring unknown sectors or reinforcing testing in recently-targeted regions. To put into perspective the significance of adaptability, we show in Fig 6 an example where an RL tester starts by targeting the same node as a degree-directed policy, but then quickly changes its behavior to also test bridging vertices. The ability to plan ahead and adapt to potential threats leads in this case to a successful

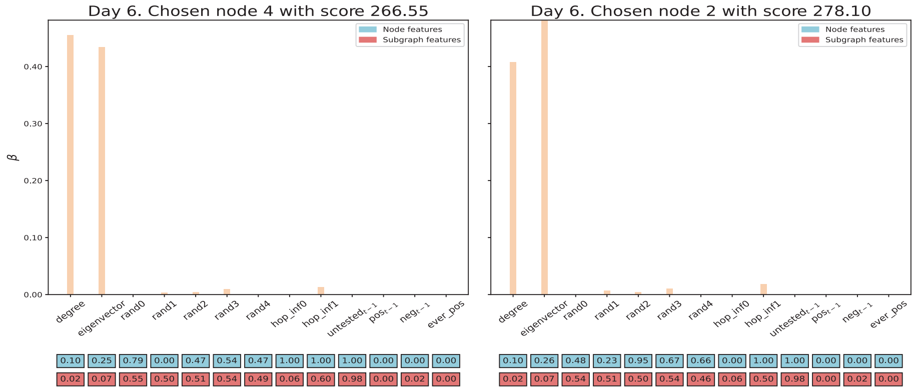


Fig. 4. Explaining early predictions on a 200 nodes network using the β importances from GraphLIME. Initially, the agent does not possess information about the epidemic state, and as such, it focuses on the centrality features. Top row displays each node's feature values, while neighborhood averages are shown underneath.

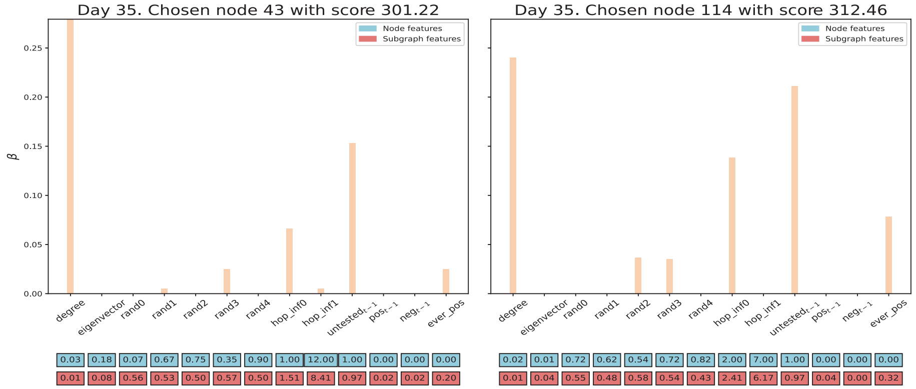
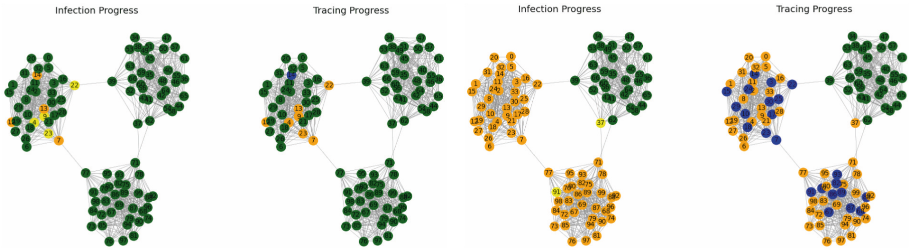
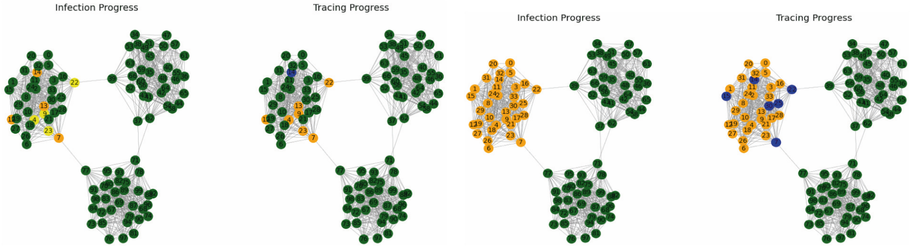


Fig. 5. Explaining later predictions on a 200 nodes network using the β importances from GraphLIME. During the later stages of an outbreak, the agent shifts its focus towards the epidemic state features, like the previously untested and positive flags, or the number of infected neighbors. Numbers in the the first row represent each node's feature values, while the second row displays the neighborhood averages.

containment of the pathogen to the first cluster, while the degree agent is unable to stop the infection of every community. We note the RL infers that bridges are important transmission vehicles in spite of never computing the time-consuming betweenness centralities (also see Appendix A.1). Considering the promising results exhibited by our RL policy, we hypothesize that useful patterns emerge within the ranking model's hidden states h_i. To verify our assumption, we plot t-SNE mappings and dendrograms for these embeddings across different days (refer to Fig A6). The detected positives (colored in blue) have a tendency to be grouped together, while new infections (red) get pushed to a handful of clusters within the same region. Such visualizations could be used for scrutinizing the actions of an agent or deriving effective community-wide health interventions.



(a) Degree w/ R: a high degree node is targeted. **(b)** Degree w/ R: all communities get infected.



(c) RL: a high degree node is targeted. **(d)** RL: only first community stays infected.

Fig. 6. Visualization of the spread for the Degree w/R and the RL agents. This corresponds to a stochastic-block network [37] with three communities. Susceptibles are green, exposed yellow, infectious orange, and detected blue. In the first day, the two policies are identical, but later on the RL agent preferentially targets the bridges. (Color figure online)

5 Conclusion and Future Work

In this study, we show how policies for controlling an epidemic through testing and tracing in a resource-limited environment can be learned using expressive graph neural networks that can integrate both local and long range infection dynamics. Across many different scenarios, a policy inferred by a reinforcement learning agent outperforms a wide range of ad-hoc rules drawing from the connectivity properties of the underlying interaction graph, achieving containment rates of up to 15% higher than degree-based solutions with recollection, and more than 50% higher than random samplers. Interestingly, our agent also exhibits strong transferability, with one model trained on small preferential attachment networks being able to control the viral diffusion on several graphs of tens of thousands of vertices and diverse linkage patterns. While building on previous efforts [65], we explore the role of contact tracing, compare different ways of modelling the infection spread (multi-mean-field versus individual agent-based), and scrutinize a varied set of heuristics. Exploring further epidemic configurations and assessing the proposed test and trace framework on real region-level data would constitute natural extensions to this work.

Additionally, we demonstrate how orderings derived by the deep learning model can be interpreted using the node features, as well as propose visualization strategies for the cluster structures that arise in the latent space of the ranking module. We believe future work could expand on the aforementioned ideas to derive more effective public health interventions and decision-making appraisals.

Acknowledgements. We thank Dr Eli Meirom for correspondence clarifying certain aspects from their study. MN was funded by EPSRC grant EP/S000356/1 Artificial and Augmented Intelligence for Automated Scientific Discovery. AR was funded by the EPSRC via a scholarship from the University of Southampton.

A Appendix

A.1 Performance Analysis

We compare the mean total elapsed time for running epidemics using each of our testing agents in Table A1. These results corresponds to the wall clock time recorded on an average Windows machine equipped with an Intel i7-7700 CPU, an NVIDIA RTX 3060 GPU and 32GB of random access memory.

Table A1. Average wall clock time per epidemic during evaluation. Configuration: Barabási-Albert networks of 2000 nodes, an average degree of approximately 3, and a daily testing budget of $k = 2$.

Agents	Wall time (s)	Agents	Wall time (s)
Random	1.12	Acquaintance (w/R)	1.12
Degree	3.23	Degree (w/R)	3.19
Closeness (w/R)	787.33	Betweenness (w/R)	1176.32
Eigenvector (w/R)	7.8	Pagerank (w/R)	6.39
Neighborhood (w/R)	1.49	RL/SL	15.92

A.2 Epidemic Modelling

All our epidemic models rely on the SEIR compartmental formulation, but the diffusion process remains bound by the interaction network configuration.

The multi-site mean-field models considered in this work rely on exponential waiting times sampled via Gillespie's algorithm to obtain subsequent events, with the state transition probabilities defined as follows:

$$\begin{aligned} p(S \rightarrow E) &= b\, w_j \triangle t \\ p(E \rightarrow I) &= e^{-1} \triangle t \\ p(I \rightarrow R) &= \rho \triangle t, \end{aligned} \tag{2}$$

where b is the base transmission rate, w_j are time-dependent edge weights, e is the exposed state duration, ρ is the recovery rate, while $\triangle t$ is a time interval.

In contrast, the agent-based model loops through every node i at every time step, executing the appropriate transition events when one of the normally-distributed samples (d_i or r_i) decreases to 0. Concurrently, every edge j is visited to check whether an infection event occurs over that connection, according to the transmission probability defined in Eq 2.

A.3 Algorithmic Details for the Proximal Policy Optimization

We start by reminding the reader about some general reinforcement learning quantities and relations:

$$\begin{gathered} \hat{A}_t^{(\gamma,0)}(a_t;\theta) = \delta_t^{\gamma}(\theta) = R_t + \gamma V(s_{t+1};\theta_T) - V(s_t;\theta) \\ \hat{A}_t^{(\gamma,1)}(a_t;\theta) = G_t^{\gamma} - V(s_t;\theta) \\ \hat{A}_t^{(\gamma,\lambda)}(a_t;\theta) = \sum_{l=0}^{T} (\gamma\lambda)^l \delta_{t+l}^{\gamma}(\theta) \\ r_t^{ORIG}(\theta) = \frac{\pi(a_t|s_t;\theta)}{\pi(a_t|s_t;\theta_k)} \quad r_t^{SARSA}(\theta) = \frac{\pi(a_t|s_{t+1};\theta)}{\pi(a_t|s_t;\theta_k)} \end{gathered} \tag{3}$$

In Eq 3, R_t is the reward obtained by the agent after taking action $a_t \sim \pi(a|s_t;\theta_k)$ and transitioning from state s_t to s_{t+1}. The value of a given state s is approximated using a neural network $V(s,\theta)$, which, together with R_t and the discount factor γ, determines the TD-error δ^{γ}; θ_k parameterizes the acting policy, θ_T is a delayed state of θ_k that parameterizes the regression target in online learning [69], $r_t^{ORIG}(\theta)$ denotes the ratio between a policy parameterized by a given θ and the acting policy, while $r_t^{SARSA}(\theta)$ represents an alternative formulation for the latter that replaces the numerator with the policy of θ evaluated at the next state s_{t+1}. Finally, the $\hat{A}_t^{(\gamma,\lambda)}$ terms represent different forms of the advantage function, as given by [87], with the special cases $\lambda = 0$, when the advantage is equal to the TD-error, and $\lambda = 1$, when the minuend of the RHS equation is the discounted return of the episode, G_t.

We rewrite the Proximal Policy Optimization (PPO) equations in terms of the quantities above in Eq 4, where $\mathcal{E}$, c_1 and c_2 are hyperparameters, $clip(.)$ is a function that clips its argument to the specified range, $transform(.)$ is a function that modifies the gradient descent update according to a specific optimizer (e.g. Adam [47]), while $\mathcal{H}_t(\theta)$ is an entropy regularizer [31]. In contrast to the original formulation, Eq 5 describes our proposed modification of PPO to allow for optimizing the objective in a memory-efficient online manner. In particular, we rewrite the loss terms using the one-step advantage function $\hat{A}_t^{(\gamma,0)}(a_t;\theta)$, and introduce an intermediary operation that accumulates the gradients of our modified loss using a unified eligibility trace [100], in a similar fashion to the methodology employed by [50], obtaining a backward-view approximation of the generalized advantage estimate $\hat{A}_t^{(\gamma,\lambda)}$ in the process [87]. We note that, by setting $r_t = r_t^{SARSA}$, we can eliminate the requirement of storing s_t in memory for the subsequent timestamp, while retaining the benefits of ratio clipping. This works well empirically since major shifts between s_{t+1} and s_t are not common in our environment. Based on previous work and our own assessment, we set $\gamma = 0.99$, $\lambda = 0.97$, $\mathcal{E} = 0.2$, $c_1 = 0.5$, $c_2 = 0.01$, and update the target value network every 5 episodes across all our experiments.

$$\begin{gathered}
\mathcal{L}_t^{CLIP}(\theta) = \min[r_t(\theta)\hat{A}_t^{(\gamma,\lambda)}(a_t;\theta), \mathrm{clip}(r_t(\theta), 1-\mathcal{E}, 1+\mathcal{E})\hat{A}_t^{(\gamma,\lambda)}(a_t;\theta)] \\
\mathcal{L}_t^{VF}(\theta) = [\hat{A}_t^{(\gamma,1)}(a_t;\theta)]^2 \quad \mathcal{H}_t(\theta) = -\sum_{a\in A}\pi(a|s_t;\theta)\log\pi(a|s_t;\theta) \\
\mathcal{L}_t^{PPO}(\theta) = \mathbb{E}_t[-\mathcal{L}_t^{CLIP}(\theta) + c_1\mathcal{L}_t^{VF}(\theta) - c_2\mathcal{H}_t(\theta))] \\
\theta_{k+1} = \arg\min_{\theta}\mathcal{L}_t^{PPO}(\theta)
\end{gathered} \tag{4}$$

$$\begin{gathered}
\mathcal{L}_t^{OCLIP}(\theta) = \min[r_t(\theta)\hat{A}_t^{(\gamma,0)}(a_t;\theta), \mathrm{clip}(r_t(\theta), 1-\mathcal{E}, 1+\mathcal{E})\hat{A}_t^{(\gamma,0)}(a_t;\theta)] \\
\mathcal{L}_t^{OVF}(\theta) = [\hat{A}_t^{(\gamma,0)}(a_t;\theta)]^2 \quad \mathcal{H}_t(\theta) = -\sum_{a\in A}\pi(a|s_t;\theta)\log\pi(a|s_t;\theta) \\
\mathcal{L}_t^{OPPO}(\theta) = -\mathcal{L}_t^{OCLIP}(\theta) + c_1\mathcal{L}_t^{OVF}(\theta) - c_2\mathcal{H}_t(\theta) \\
E_t = \gamma\lambda E_{t-1} + \frac{\nabla_{\theta_k}\mathcal{L}_t^{OPPO}(\theta_k)}{s}\text{, with } s = \delta_t^{\gamma}(\theta_k) \text{ or } s = 1 \\
\Delta\theta_k = \mathrm{transform}(\delta_t^{\gamma}(\theta_k)E_t) \\
\theta_{k+1} = \theta_k - \Delta\theta_k
\end{gathered} \tag{5}$$

A.4 Supporting Figures

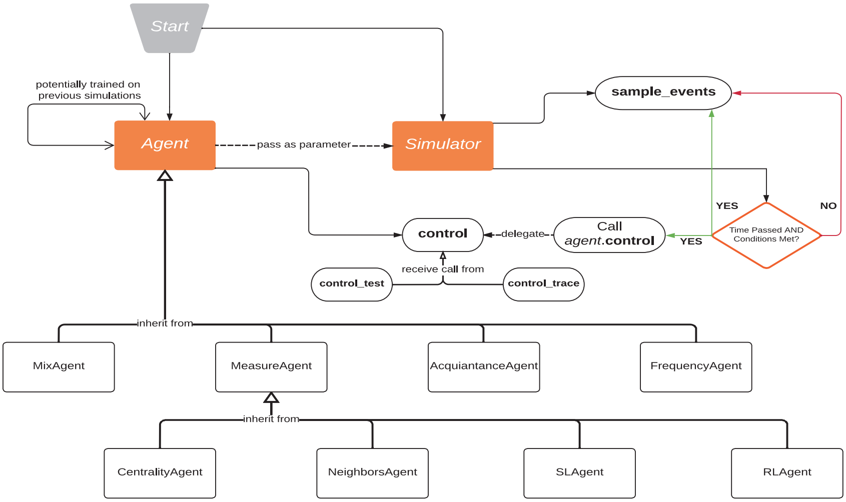


Fig. A1. Block diagram of our control framework. The *Agent* is passed as a parameter to the *Simulator*, and every time the latter samples enough events for the conditions to be met, a call to the *control(.)* method of the first is performed. The aforementioned function performs some preprocessing steps, and then calls *control_test(.)* and *control_trace(.)*, which are responsible for the actual node ranking and are specific to each type of agent. Combinations of agents can be selected with the *MixAgent*.

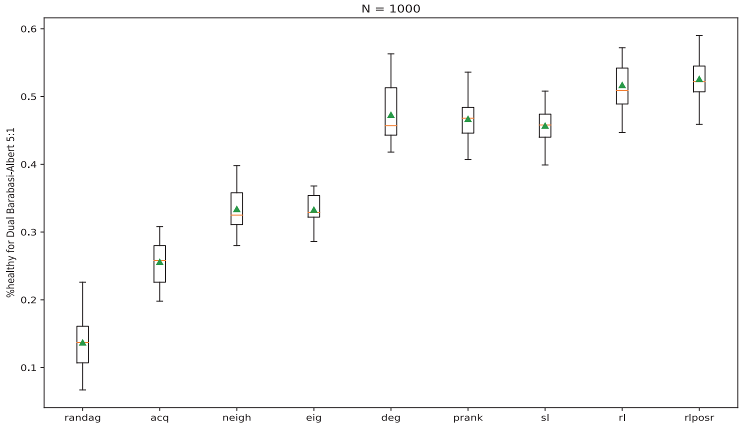


Fig. A2. Infection control performance on different network architectures of 1000 nodes and a daily testing budget of $k = 2$. Uncertainties shown as boxplots.

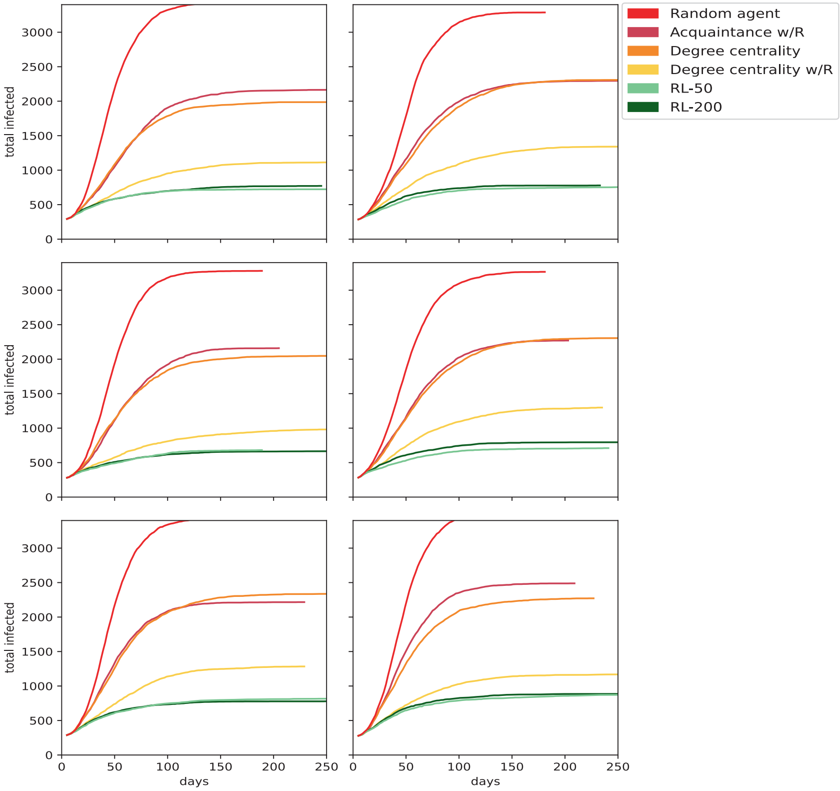


Fig. A3. Epidemic curves for different network and epidemic seeds. These correspond to multiple 5000 nodes Barabási-Albert networks featuring a mean degree of 3, with a testing budget of $k = 1\%$. Here, two versions of the RL agent are displayed: one trained for 50, and one trained for 200 episodes. The y-axis limit is set to 3200 to facilitate the comparisons, yet the random agents perform poorer than this level.

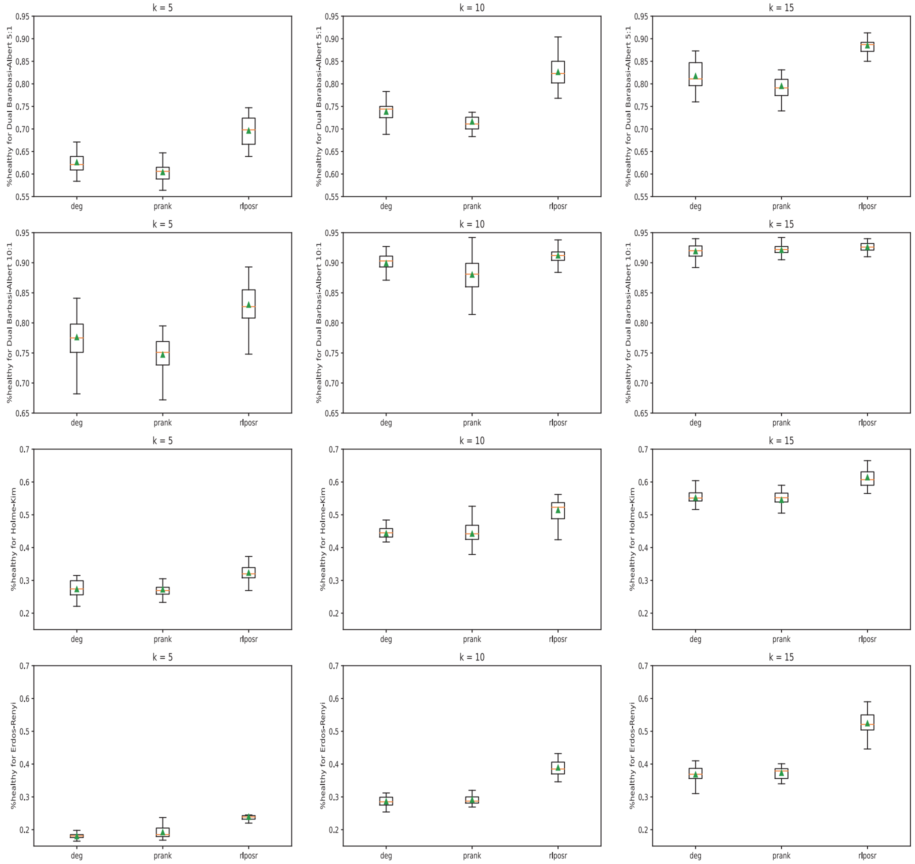


Fig. A4. Infection control performance on different static network architectures with varying budgets. The uncertainties are shown as boxplots.

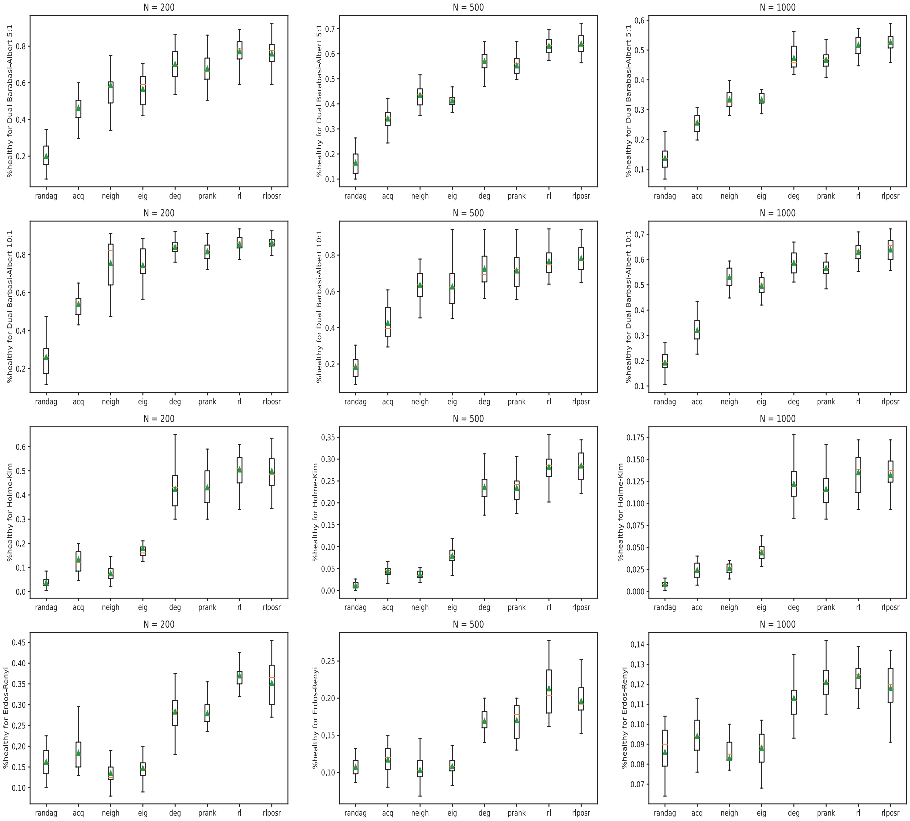


Fig. A5. Infection control performance on different static network architectures and sizes, with a budget of $k = 2$. Uncertainties are shown as boxplots.

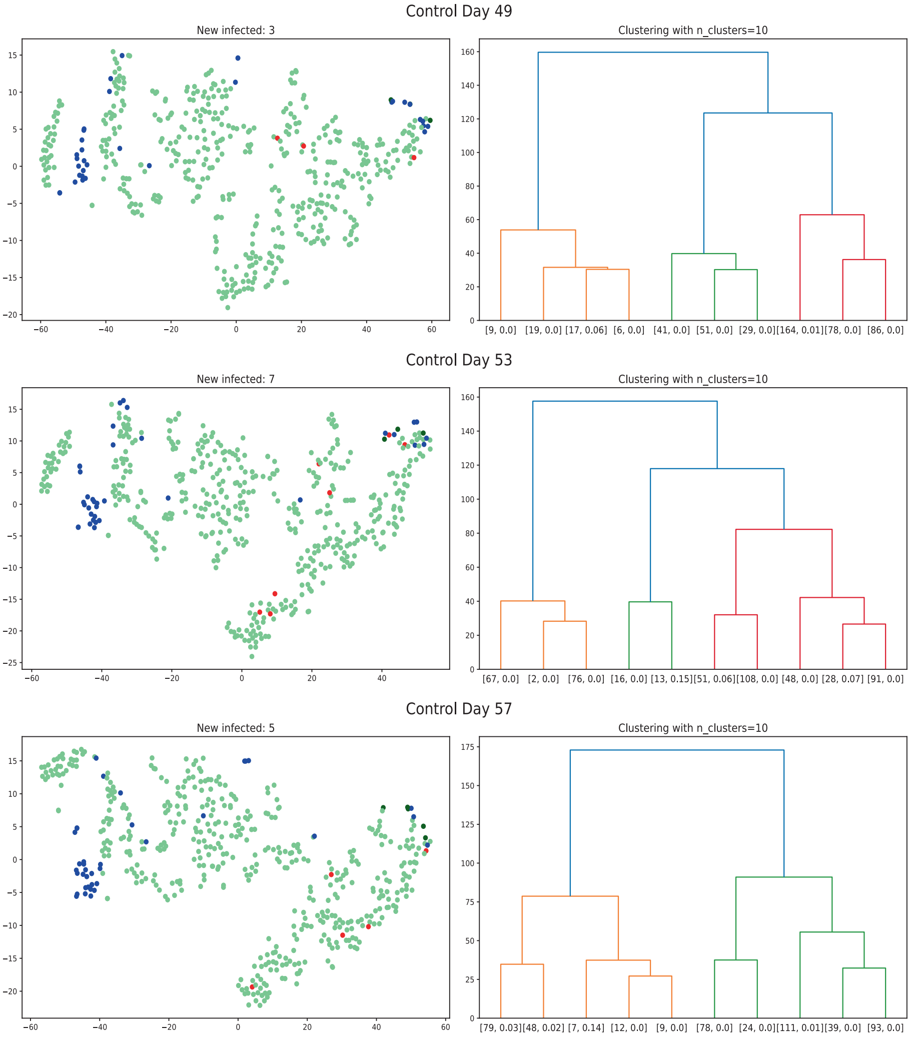


Fig. A6. t-SNE plots of the node hidden states and dendrogram corresponding to their hierarchical clustering into 10 groups. As can be observed, the agent mostly groups detected (blue) nodes in a region of the space, while the new undetected infections (red) are predicted to appear within the risk regions on the right. Recent negative results are plotted as dark green. The dendrogram on the right displays the cardinality and the infection probability associated with each cluster.

A.5 Control Framework

The logic behind our epidemic control framework in the continuous-time simulation scenario is outlined in Algorithm 1. The class hierarchy of the agents, together with their logic, can be consulted in Algorithm 2. Refer to Table A2 for details about the variables involved in these.

Table A2. Legend for the control framework pseudocode.

Name(s)	**Description(s)**
R	GNN-based ranking model (shared across all epidemics).
E_{conf}	Episode configuration. Consists of tuples mapping an episode ID (e_{id}) to its exploration-control variable e_{ϵ}.
S_{conf}	Simulation configuration. Enum that defines the maximum network, infection and event seeds, which in turn control the range of the loops over each seeded configuration.
s_{net}, s_{inf}, s_{ev}	Interaction network, infection and event seeds.
N_p, S_p, A_p	Interaction network, simulator and agent hyperparameters. A_p contains the sampling strategy st and learning rate lr.
N, S, A	Interaction network, simulator and agent main objects.
i_c, i_u	Iterators for time-discretized events: dynamic control and edge-updating interaction events.
e, t	Interaction event enum and its corresponding time value.
k_{tst}, k_{ct}	Daily budgets for testing and contact-tracing isolations.
c_{tst}, c_{ct}	Sensible candidates to rank for testing and tracing.
n_{tst}, n_{ct}	Nodes chosen by the agent for testing and tracing.
d	Boolean that determines whether the action is sampled or greedily taken from top-k ranking.
st	Sampling strategy employed by the RLAgent. This can be one of the following: 'softmax', 'escort-transform' [64], 'nvidia-explore' [65].
m	Node ranking scores computed by a specific agent.
v	Epidemic state score computed by the GNN ranking model.
B	Replay buffer for the offline RLAgent.
L	Last step information required by the online RLAgent.
a, $\log \pi_a$	Sampled action and its corresponding log of probability.
r_{t-1}	Reward of previous action taken (i.e. for action sampled and executed at time $t-1$).

Algorithm 1 Epidemic control framework

```
global variables
    $R$                                                        ▷ GNN ranking model
end global variables
procedure RUN_EPIDEMIC($E_{conf}$, $S_{conf}$, $N_p$, $S_p$, $A_p$)
    for each $(e_{id}, e_\epsilon) \in E_{conf}$ do                         ▷ Episode ID and $\epsilon$
        for $s_{net} \in \{0, \ldots, S_{conf}$.MAX_NET_SEED$\}$ do
            ▷ $N$ keeps in memory the edges over all timestamps
            $N \leftarrow$ INIT_NET($N_p$, $s_{net}$)
            for $s_{inf} \in \{0, \ldots, S_{conf}$.MAX_INF_SEED$\}$ do
                $S \leftarrow$ INIT_SIMULATOR($S_p$, $s_{inf}$, $N$)
                $A \leftarrow$ INIT_AGENT($A_p$, $R$, $e_{id}$, $e_\epsilon$)
                $i_c \leftarrow 0$                              ▷ Iterator for control timestamps
                $i_u \leftarrow 0$                          ▷ Iterator for edge-update timestamps
                for $s_{ev} \in \{0, \ldots, S_{conf}$.MAX_EVENT$\}$ do
                    $e \leftarrow S$.SAMPLE_NEXT_EVENT()
                    $t \leftarrow e$.TIME
                    $S$.RUN_EVENT($e$, $N$)
                    if $S$.SHOULD_CONTROL($N$, $t$, $i_c$) then
                        $i_c \leftarrow \lfloor t \rfloor$                                   ▷ Floor function
                        $(n_{tst}, n_{ct}) \leftarrow A$.CONTROL($N$, $i_c$)
                        $S$.UPDATE_STATES($N$, $n_{tst}$, $n_{ct}$)
                        $i_c \leftarrow i_c + 1$
                    end if
                    if $S$.SHOULD_UPDATE_EDGES($N$, $t$, $i_u$) then
                        $i_u \leftarrow \lfloor t \rfloor$
                        $N$.UPDATE_EDGES($i_u$)
                        $i_u \leftarrow i_u + 1$
                    end if
                end for
                ▷ Logging & offline parameter updates (if any)
                $A$.FINISH($N$)
            end for
        end for
    end for
end procedure
```

Algorithm 2 Control agents' hierarchy

```
struct Agent
    $k_{tst}$, $k_{ct}$    ▷ Budget for testing and contact tracing
    procedure control($N$, $i_c$)
        $c_{tst} \leftarrow$ candidates_test($N$, $i_c$)
        ▷ Calls control_both by default; can be overridden
        $n_{tst} \leftarrow$ control_test($N$, $c_{tst}$, $k_{tst}$)
        $c_{ct} \leftarrow$ candidates_trace($N$, $i_c$, $n_{tst}$)
        ▷ Calls control_both by default; can be overridden
        $n_{ct} \leftarrow$ control_trace($N$, $c_{ct}$, $k_{ct}$)
        return ($n_{tst}$, $n_{ct}$)
    end procedure
end struct
struct MeasureAgent(Agent)
    $d$    ▷ Boolean controlling if sampling or top-k ranking
    procedure control_both($N$, $c$, $k$)
        ▷ Compute score for each node in $c$; RL samples $k$ nodes
        $m \leftarrow$ compute_measures($N$, $c$, $k$)
        if $d$ then
            return $c$[argtopk($m$, $k$)]    ▷ Heap sort top-k ranking
        else
            return $m$
        end if
    end procedure
end struct
struct SLAgent(MeasureAgent)
    $lr$    ▷ Learning rate; if 0, evaluation mode is assumed
    procedure compute_measures($N$, $c$, $k$)
        if $lr > 0$ then
            ($m$, $v$) $\leftarrow$ $R$.forward($N$)    ▷ Message passing
            backprop_loss($N$, $m$)    ▷ BCE on infection status
        else
            ($m$, $v$) $\leftarrow$ $R$.forward(subgraph($N$, $c$))
        end if
        return $m$
    end procedure
end struct
struct RLAgent(MeasureAgent)
    $lr$    ▷ Learning rate; if 0, evaluation mode is assumed
    $st$    ▷ Action sampling strategy (e.g. softmax)
    $e_\epsilon$    ▷ Sampling noise (i.e. $\epsilon$-greedy, softmax temperature)
    $B$    ▷ Replay buffer; if null, conduct online learning
    $L$    ▷ Keep last step information for online learning
    procedure compute_measures($N$, $c$, $k$)
        if $lr > 0$ then
            ▷ Reward of previous action
            $R_{t-1} \leftarrow -N$.num_infected()
            ($m$, $v$) $\leftarrow$ $R$.forward($N$)    ▷ Message passing
            ($a$, $\log \pi_a$) $\leftarrow$ sample($m$, $k$, $st$, $e_\epsilon$)    ▷ Sample action
            ▷ Existence of $B$ determines training online/offline
            if $B$ is null then
                ▷ Compute online RL objective
                ▷ Using $m$, $\{s_{t-1}, a_{t-1}\} \in L$ to compute $\log \pi_{a_{t-1}}$
                backprop_loss($R_{t-1}$, $L$, $m$, $v$)
                $L$.clear()
                ▷ Add ($s_t$, $a_t$, $\log \pi_{a_t}$, $V_t$) to one-step buffer
                $L$.add($N$, $a$, $\log \pi_a$, $v$)    ▷ $\log \pi_a$ used for $r_t^{SARSA}$
            else
                ▷ Add ($R_{t-1}$, $s_t$, $a_t$, $\log \pi_{a_t}$, $V_t$) to replay buffer
                $B$.add($R_{t-1}$, $N$, $a$, $\log \pi_a$, $v$)
            end if
        else
            ($m$, $v$) $\leftarrow$ $R$.forward(subgraph($N$, $c$))
        end if
        return $m$
    end procedure
end struct
```

References

1. Abueg, M., et al.: Modeling the combined effect of digital exposure notification and non-pharmaceutical interventions on the COVID-19 epidemic in Washington state. In: medRxiv, p. 2020.08.29.20184135. Cold Spring Harbor Laboratory Press (2020). https://doi.org/10.1101/2020.08.29.20184135
2. Alon, U., Yahav, E.: On the bottleneck of graph neural networks and its practical implications. In: International Conference on Learning Representations (2022)
3. Andrews, N., et al.: COVID-19 vaccine effectiveness against the omicron (B.1.1.529) variant. New Engl. J. Med. **386**(16), 1532–1546 (2022). https://doi.org/10.1056/NEJMoa2119451
4. Bao, H., Dong, L., Wei, F.: BEiT: BERT pre-training of image transformers (2021). https://doi.org/10.48550/arXiv.2106.08254
5. Bastani, H., et al.: Efficient and targeted COVID-19 border testing via reinforcement learning. Nature **599**(7883), 108–113 (2021). https://doi.org/10.1038/s41586-021-04014-z, https://www.nature.com/articles/s41586-021-04014-z
6. Beaini, D., Passaro, S., Létourneau, V., Hamilton, W.L., Corso, G., Liò, P.: Directional Graph Networks (2021). https://doi.org/10.48550/arXiv.2010.02863
7. Bello, I., Pham, H., Le, Q.V., Norouzi, M., Bengio, S.: Neural combinatorial optimization with reinforcement learning (2017). https://doi.org/10.48550/arXiv.1611.09940
8. Bodnar, C., Di Giovanni, F., Chamberlain, B.P., Liò, P., Bronstein, M.M.: Neural sheaf diffusion: a topological perspective on heterophily and oversmoothing in GNNs (2022). https://doi.org/10.48550/arXiv.2202.04579
9. Braha, D., Bar-Yam, Y.: From centrality to temporary fame: dynamic centrality in complex networks. Complexity **12**(2), 59–63 (2006). https://doi.org/10.1002/cplx.20156
10. Brody, S., Alon, U., Yahav, E.: How attentive are graph attention networks? (2022). https://doi.org/10.48550/arXiv.2105.14491
11. Bronstein, M.: Deep learning on graphs: successes, challenges, and next steps (2022). https://towardsdatascience.com/deep-learning-on-graphs-successes-challenges-and-next-steps-7d9ec220ba8
12. Bruxvoort, K.J., et al.: Effectiveness of mRNA-1273 against delta, mu, and other emerging variants of SARS-CoV-2: test negative case-control study. BMJ **375**, e068848 (2021). https://doi.org/10.1136/bmj-2021-068848
13. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 2009, p. 199. ACM Press, Paris (2009). https://doi.org/10.1145/1557019.1557047
14. Chung, F., Horn, P., Tsiatas, A.: Distributing antidote using PageRank vectors. Internet Math. **6**(2), 237–254 (2009). https://doi.org/10.1080/15427951.2009.10129184
15. Clair, R., Gordon, M., Kroon, M., Reilly, C.: The effects of social isolation on well-being and life satisfaction during pandemic. Humanit. Soc. Sci. Commun. **8**(1), 1–6 (2021). https://doi.org/10.1057/s41599-021-00710-3
16. Cohen, R., Havlin, S., ben-Avraham, D.: Efficient immunization strategies for computer networks and populations. Phys. Rev. Lett. **91**(24), 247901 (2003). https://doi.org/10.1103/PhysRevLett.91.247901
17. Dai, H., Khalil, E.B., Zhang, Y., Dilkina, B., Song, L.: Learning combinatorial optimization algorithms over graphs (2018)

18. Davis, E.L., et al.: Contact tracing is an imperfect tool for controlling COVID-19 transmission and relies on population adherence. Nat. Commun. **12**(1), 5412 (2021). https://doi.org/10.1038/s41467-021-25531-5
19. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs] (2019)
20. Di Domenico, L., Pullano, G., Sabbatini, C.E., Boëlle, P.Y., Colizza, V.: Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies. BMC Med. **18**(1), 240 (2020). https://doi.org/10.1186/s12916-020-01698-4
21. Dighe, A., et al.: Response to COVID-19 in South Korea and implications for lifting stringent interventions. BMC Med. **18**(1), 321 (2020). https://doi.org/10.1186/s12916-020-01791-8
22. Erdös, P., Rényi, A.: On random graphs I. Publicationes Mathematicae Debrecen **6**, 290 (1959)
23. Farrahi, K., Emonet, R., Cebrian, M.: Epidemic contact tracing via communication traces. PLoS ONE **9**(5), e95133 (2014). https://doi.org/10.1371/journal.pone.0095133
24. Ferdinands, J.M.: Waning 2-Dose and 3-dose effectiveness of mrna vaccines against COVID-19–associated emergency department and urgent care encounters and hospitalizations among adults during periods of delta and omicron variant predominance—VISION network, 10 States, August 2021–January 2022. MMWR Morbidity Mortality Weekly Rep. **71** (2022). https://doi.org/10.15585/mmwr.mm7107e2
25. Ferguson, N., et al.: Report 9: impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. Technical report, Imperial College London (2020). https://doi.org/10.25561/77482
26. Ferretti, L., et al.: Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science **368** (2020). https://doi.org/10.1126/science.abb6936
27. Fung, V., Zhang, J., Juarez, E., Sumpter, B.: Benchmarking graph neural networks for materials chemistry. NPJ Comput. Mater. **7**, 84 (2021). https://doi.org/10.1038/s41524-021-00554-0
28. Gillespie, D.T.: Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. **81**(25), 2340–2361 (1977). https://doi.org/10.1021/j100540a008
29. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E.: Neural message passing for quantum chemistry (2017). https://doi.org/10.48550/arXiv.1704.01212
30. Gori, M., Monfardini, G., Scarselli, F.: A new model for earning in graph domains. In: Proceedings of the International Joint Conference on Neural Networks, vol. 2, pp. 729–734 (2005). https://doi.org/10.1109/IJCNN.2005.1555942
31. Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor (2018)
32. Hamilton, W.L., Ying, R., Leskovec, J.: Inductive representation learning on large graphs. arXiv:1706.02216 [cs, stat] (2018)
33. He, J., et al.: Deep reinforcement learning with a combinatorial action space for predicting popular reddit threads. In: EMNLP (2019)
34. Henley, J.: COVID surges across Europe as experts warn not let guard down. The Guardian (2022). https://www.theguardian.com/world/2022/jun/21/covid-surges-europe-ba4-ba5-cases
35. Hinch, R., et al.: Effective configurations of a digital contact tracing app: a report to NHSX. Technical report (2020)

36. Hoang, N., Maehara, T.: Revisiting graph neural networks: all we have is low-pass filters. arXiv:1905.09550 [cs, math, stat] (2019)
37. Holland, P.W., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: first steps. Soc. Netw. **5**(2), 109–137 (1983). https://doi.org/10.1016/0378-8733(83)90021-7
38. Holme, P., Kim, B.J.: Growing scale-free networks with tunable clustering. Phys. Rev. E **65**(2), 026107 (2002). https://doi.org/10.1103/PhysRevE.65.026107
39. Huang, Q., Yamada, M., Tian, Y., Singh, D., Yin, D., Chang, Y.: GraphLIME: local interpretable model explanations for graph neural networks (2020). https://doi.org/10.48550/arXiv.2001.06216
40. Huerta, R., Tsimring, L.S.: Contact tracing and epidemics control in social networks. Phys. Rev. E **66**(5), 056115 (2002). https://doi.org/10.1103/PhysRevE.66.056115
41. Jhun, B.: Effective vaccination strategy using graph neural network ansatz (2021). https://doi.org/10.48550/arXiv.2111.00920
42. Joffe, A.R.: COVID-19: rethinking the lockdown groupthink. Front. Public Health **9** (2021)
43. Joshi, C.K., Laurent, T., Bresson, X.: An efficient graph convolutional network technique for the travelling salesman problem (2019)
44. Kapoor, A., et al.: Examining COVID-19 forecasting using spatio-temporal graph neural networks. arXiv:2007.03113 [cs] (2020)
45. Kermack, W.O., McKendrick, A.G., Walker, G.T.: A contribution to the mathematical theory of epidemics. Proc. Roy. Soc. London Ser. A Containing Pap. Math. Phys. Character **115**(772), 700–721 (1927). https://doi.org/10.1098/rspa.1927.0118
46. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: a survey. ACM Comput. Surv. 3505244 (2022). https://doi.org/10.1145/3505244
47. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2017). https://doi.org/10.48550/arXiv.1412.6980
48. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017. Conference Track Proceedings (2017). OpenReview.net
49. Kiran, B.R., et al.: Deep reinforcement learning for autonomous driving: a survey. IEEE Trans. Intell. Transp. Syst. **23**(6), 4909–4926 (2022). https://doi.org/10.1109/TITS.2021.3054625
50. Kobayashi, T.: Adaptive and multiple time-scale eligibility traces for online deep reinforcement learning. Robot. Auton. Syst. **151**, 104019 (2022). https://doi.org/10.1016/j.robot.2021.104019
51. Kojaku, S., Hébert-Dufresne, L., Mones, E., Lehmann, S., Ahn, Y.Y.: The effectiveness of backward contact tracing in networks. Nat. Phys. **17**(5), 652–658 (2021). https://doi.org/10.1038/s41567-021-01187-2
52. Konda, V., Tsitsiklis, J.: Actor-critic algorithms. In: Advances in Neural Information Processing Systems, vol. 12. MIT Press (1999)
53. Kool, W., van Hoof, H., Welling, M.: Attention, learn to solve routing problems! (2019). https://doi.org/10.48550/arXiv.1803.08475
54. Lazaridis, A., Fachantidis, A., Vlahavas, I.: Deep reinforcement learning: a state-of-the-art walkthrough. J. Artif. Intell. Res. **69**, 1421–1471 (2020). https://doi.org/10.1613/jair.1.12412

55. Leung, K., Wu, J.T.: Managing waning vaccine protection against SARS-CoV-2 variants. Lancet **399**(10319), 2–3 (2022). https://doi.org/10.1016/S0140-6736(21)02841-5
56. Lillicrap, T.P., et al.: Continuous control with deep reinforcement learning. Paper presented at 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico (2016)
57. Liu, D., Jing, Y., Zhao, J., Wang, W., Song, G.: A fast and efficient algorithm for mining top-k nodes in complex networks. Sci. Rep. **7**(1), 43330 (2017). https://doi.org/10.1038/srep43330
58. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions (2017). https://doi.org/10.48550/arXiv.1705.07874
59. Madan, A., Cebrian, M., Moturu, S., Farrahi, K., Pentland, A.S.: Sensing the "health state" of a community. IEEE Pervasive Comput. **11**(4), 36–45 (2012). https://doi.org/10.1109/MPRV.2011.79
60. Martinez-Garcia, M., Sansano-Sansano, E., Castillo-Hornero, A., Femenia, R., Roomp, K., Oliver, N.: Social isolation during the COVID-19 pandemic in Spain: a population study (2022). https://doi.org/10.1101/2022.01.22.22269682
61. Mason, R., Allegretti, A., Devlin, H., Sample, I.: UK treasury pushes to end most free Covid testing despite experts' warnings. The Guardian (2022)
62. Masuda, N.: Immunization of networks with community structure. New J. Phys. **11**(12), 123018 (2009). https://doi.org/10.1088/1367-2630/11/12/123018
63. Matrajt, L., Leung, T.: Evaluating the effectiveness of social distancing interventions to delay or flatten the epidemic curve of coronavirus disease. Emerg. Infect. Dis. **26**(8), 1740–1748 (2020). https://doi.org/10.3201/eid2608.201093
64. Mei, J., Xiao, C., Dai, B., Li, L., Szepesvari, C., Schuurmans, D.: Escaping the gravitational pull of softmax. In: Advances in Neural Information Processing Systems, vol. 33, pp. 21130–21140. Curran Associates, Inc. (2020)
65. Meirom, E., Maron, H., Mannor, S., Chechik, G.: Controlling graph dynamics with reinforcement learning and graph neural networks. In: Proceedings of the 38th International Conference on Machine Learning, pp. 7565–7577. PMLR (2021)
66. Meirom, E., Milling, C., Caramanis, C., Mannor, S., Shakkottai, S., Orda, A.: Localized epidemic detection in networks with overwhelming noise. ACM SIGMETRICS Perform. Eval. Rev. **43**(1), 441–442 (2015). https://doi.org/10.1145/2796314.2745883
67. Mercer, T.R., Salit, M.: Testing at scale during the COVID-19 pandemic. Nat. Rev. Genet. **22**(7), 415–426 (2021). https://doi.org/10.1038/s41576-021-00360-w
68. Miller, J.C., Hyman, J.M.: Effective vaccination strategies for realistic social networks. Phys. A **386**(2), 780–785 (2007). https://doi.org/10.1016/j.physa.2007.08.054
69. Mnih, V., et al.: Playing atari with deep reinforcement learning (2013). https://doi.org/10.48550/arXiv.1312.5602
70. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015). https://doi.org/10.1038/nature14236
71. Morris, C., et al.: Weisfeiler and leman go neural: higher-order graph neural networks. arXiv:1810.02244 [cs, stat] (2020)
72. Moshiri, N.: The dual-Barabási-Albert model (2018)
73. Murata, T., Koga, H.: Extended methods for influence maximization in dynamic networks. Comput. Soc. Netw. **5**(1), 1–21 (2018). https://doi.org/10.1186/s40649-018-0056-8
74. Oono, K., Suzuki, T.: Graph neural networks exponentially lose expressive power for node classification. arXiv:1905.10947 [cs, stat] (2021)

75. Panagopoulos, G., Nikolentzos, G., Vazirgiannis, M.: Transfer graph neural networks for pandemic forecasting. arXiv:2009.08388 [cs, stat] (2021)
76. Pandit, J.A., Radin, J.M., Quer, G., Topol, E.J.: Smartphone apps in the COVID-19 pandemic. Nat. Biotechnol. **40**(7), 1013–1022 (2022). https://doi.org/10.1038/s41587-022-01350-x
77. Preciado, V.M., Zargham, M., Enyioha, C., Jadbabaie, A., Pappas, G.J.: Optimal resource allocation for network protection against spreading processes. IEEE Trans. Control Netw. Syst. **1**(1), 99–108 (2014). https://doi.org/10.1109/TCNS.2014.2310911
78. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming, 1st edn. Wiley, USA (1994)
79. Rayner, D.C., Sturtevant, N.R., Bowling, M.: Subset selection of search heuristics. In: IJCAI (2019)
80. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?": explaining the predictions of any classifier (2016). https://doi.org/10.48550/arXiv.1602.04938
81. Rimmer, A.: Sixty seconds on . . . the pingdemic. BMJ **374**, n1822 (2021). https://doi.org/10.1136/bmj.n1822
82. Rummery, G., Niranjan, M.: On-line Q-learning using connectionist systems. Technical report CUED/F-INFENG/TR 166 (1994)
83. Rusu, A., Farrahi, K., Emonet, R.: Modelling digital and manual contact tracing for COVID-19 Are low uptakes and missed contacts deal-breakers? Preprint. Epidemiology (2021). https://doi.org/10.1101/2021.04.29.21256307
84. Salathé, M., Jones, J.H.: Dynamics and control of diseases in networks with community structure. PLOS Comput. Biol. **6**(4), e1000736 (2010). https://doi.org/10.1371/journal.pcbi.1000736, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000736
85. Sato, R., Yamada, M., Kashima, H.: Random features strengthen graph neural networks (2021). https://doi.org/10.48550/arXiv.2002.03155
86. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE Trans. Neural Netw. **20**(1), 61–80 (2009). https://doi.org/10.1109/TNN.2008.2005605
87. Schulman, J., Moritz, P., Levine, S., Jordan, M., Abbeel, P.: High-dimensional continuous control using generalized advantage estimation (2018). https://doi.org/10.48550/arXiv.1506.02438
88. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv:1707.06347 [cs] (2017)
89. Serafino, M., et al.: Digital contact tracing and network theory to stop the spread of COVID-19 using big-data on human mobility geolocalization. PLOS Comput. Biol. **18**(4), e1009865 (2022). https://doi.org/10.1371/journal.pcbi.1009865
90. Shah, C., et al.: Finding patient zero: learning contagion source with graph neural networks (2020)
91. Sigal, A.: Milder disease with Omicron: is it the virus or the pre-existing immunity? Nat. Rev. Immunol. **22**(2), 69–71 (2022). https://doi.org/10.1038/s41577-022-00678-4
92. Silver, D., et al.: Mastering the game of Go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016). https://doi.org/10.1038/nature16961
93. Silver, D., et al.: Mastering chess and shogi by self-play with a general reinforcement learning algorithm (2017). https://doi.org/10.48550/arXiv.1712.01815
94. Smith, J.: Demand for Covid vaccines falls amid waning appetite for booster shots. Financial Times (2022). https://www.ft.com/content/9ac9f8fc-1ab3-4cb2-81bf-259ba612f600

95. Smith, R.L., et al.: Longitudinal assessment of diagnostic test performance over the course of acute SARS-CoV-2 infection. J. Infect. Dis. **224**(6), 976–982 (2021). https://doi.org/10.1093/infdis/jiab337
96. Song, H., et al.: Solving continual combinatorial selection via deep reinforcement learning. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, pp. 3467–3474 (2019). https://doi.org/10.24963/ijcai.2019/481
97. Su, Z., Cheshmehzangi, A., McDonnell, D., da Veiga, C.P., Xiang, Y.T.: Mind the "Vaccine Fatigue". Front. Immunol. **13** (2022)
98. Sukumar, S.R., Nutaro, J.J.: Agent-based vs. equation-based epidemiological models: a model selection case study. In: 2012 ASE/IEEE International Conference on BioMedical Computing (BioMedCom), pp. 74–79 (2012). https://doi.org/10.1109/BioMedCom.2012.19
99. Sutton, R.S.: Learning to predict by the methods of temporal differences. Mach. Learn. **3**(1), 9–44 (1988). https://doi.org/10.1007/BF00115009
100. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. Adaptive Computation and Machine Learning Series, 2nd edn. The MIT Press, Cambridge (2018)
101. Tian, S., Mo, S., Wang, L., Peng, Z.: Deep reinforcement learning-based approach to tackle topic-aware influence maximization. Data Sci. Eng. **5**(1), 1–11 (2020). https://doi.org/10.1007/s41019-020-00117-1
102. Tomy, A., Razzanelli, M., Di Lauro, F., Rus, D., Della Santina, C.: Estimating the state of epidemics spreading with graph neural networks. Nonlinear Dyn. **109**(1), 249–263 (2022). https://doi.org/10.1007/s11071-021-07160-1
103. van Hasselt, H., Madjiheurem, S., Hessel, M., Silver, D., Barreto, A., Borsa, D.: Expected eligibility traces. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 9997–10005 (2021). https://doi.org/10.1609/aaai.v35i11.17200
104. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
105. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. arXiv:1710.10903 [cs, stat] (2018)
106. Watkins, C.: Learning from delayed rewards (1989)
107. Wymant, C., et al.: The epidemiological impact of the NHS COVID-19 app. Nature **594**(7863), 408–412 (2021). https://doi.org/10.1038/s41586-021-03606-z
108. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? arXiv:1810.00826 [Cs, Stat] (2019)
109. Yamada, M., Jitkrittum, W., Sigal, L., Xing, E.P., Sugiyama, M.: High-dimensional feature selection by feature-wise kernelized lasso. Neural Comput. **26**(1), 185–207 (2014). https://doi.org/10.1162/NECO_a_00537
110. Zhou, J., et al.: Graph neural networks: a review of methods and applications. AI Open **1**, 57–81 (2020). https://doi.org/10.1016/j.aiopen.2021.01.001

Connecting Self-reported COVID-19 Needs with Social Determinants of Health

Jessica A. Pater[1](✉), Tammy Toscos[1], Mindy Flanagan[1], Michelle Drouin[1], Deborah McMahan[2], Meg Distler[3], Patti Hayes[4], and Nelson Peters[5]

[1] Parkview Research Center, Fort Wayne, IN, USA
`Jessica.Pater@parkview.com`
[2] Allen County Department of Public Health, Fort Wayne, IN, USA
[3] St. Joseph Community Health Foundation, Fort Wayne, IN, USA
[4] AWS Foundation, Fort Wayne, IN, USA
[5] Allen County Government, Fort Wayne, IN, USA

Abstract. COVID-19 rapidly challenged and changed our understanding of what needs were unmet in the community and the reality of how stable communities were with respect to basic daily needs like transportation, access to medications, how financial reserves. In this study, we report on a set of hyper-local community-based surveys (N = 44796; N = 1039) developed by stakeholders from across the community using a social determinants of health lens to rapidly measure these evolving needs. Findings were stratified across a financial sustainability measure and focused on understanding where people would and were looking for support for medication and healthcare needs as well as the basic life necessities of food, water, utilities, and shelter. Survey results were shared with health system and community leaders as well as elected officials to support real-time data-driven decision making within our local community as needs rapidly evolved.

Keywords: COVID-19 · Social Determinants of Health (SDOH) · community survey · community partnerships

1 Introduction

The novel coronavirus disease 2019 (COVID-19) was a world-wide pandemic that struck rapidly causing approximately 56,498,113 infections and 1,345,205 deaths within the first months of the spread, from March to November 2020. Identifying the social impacts of COVID-19 is essential in understanding the totality of the pandemic. Research into social impacts span from public perceptions [1] to mental health effects [2] to disparity related to social determinants of health (SDOH), such as racial/ethnic based differences [4,5], socioeconomic indicators [6], and the need for targeted response to support communities most vulnerable to complications of COVID-19 infection [7]. Emerging data within the United

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 207–225, 2023.
https://doi.org/10.1007/978-3-031-34586-9_15

States shows that using social determinants of health as a lens to community-level impacts of COVID-19 is useful. These inquiries have largely been focused on treatment of and susceptibility to the disease among specific communities. One example of these efforts comes from Chin et al., who created a SDOH-focused vulnerability map of U.S. counties, highlighting social determinants that might increase or decrease its residents' risk of contracting the disease (e.g., age, population density, poverty, job insecurity, and health insurance) per county [7]. This geocoding map extends research demonstrating that particular SDOH factors relate to hospitalization and mortality. For example, a study conducted in New York City showed that the highest rates of COVID-19 hospitalizations and deaths were found in neighborhoods with the highest rates of poverty [8]. Additionally, COVID-19 has been found to disproportionally impact minority communities in both infection and mortality rates [9,10], and thus the allocations of resources and support should include this consideration [11].

Although most of COVID-19 studies address SDOH factors from disease vulnerability, treatment, or mortality perspective, the impacts of COVID on communities is broader than just the impact of the disease itself. Indeed, some disparities in SDOH factors that lead to illness vulnerability (e.g., financial instability and access to health care and support services), might be exacerbated as communities were coping with the pandemic and its associated mandates (e.g. limited hours of operation for businesses). The movement restrictions imposed by COVID-19 mandates alone were related to severe economic costs among both regional and national governments and their vulnerable citizens [12]. Hence, when communities chose to respond to COVID-19 and health disparities, generally, a coordinated effort between municipal, philanthropic, and public and private organizations was essential to tackle the myriad effects the pandemic or other public health threats might bring. In line with models which address health disparities through grants and coordinated efforts among community organizations [13], this research study was developed to understand the impacts of COVID-19 and consequences of public health restrictions put in place to control the spread of the infection within a localized context in order to support decision-making for resource allocation and concentrated community aid.

A collaboration in Northeast Indiana was formed between the Allen County Health Department, Parkview Health System and local business, city and philanthropy leaders. Stakeholders met in early March 2020 to develop a set of community surveys. These surveys focused on three key foci for social determinants of health as defined in Healthy People 2020: 1) economic stability (employment, food insecurity, housing instability, poverty), 2) health and healthcare (access to health care, access to primary care, health literacy), and 3) neighborhood and built environment (access to local support services, crime and violence, environmental conditions). Due to the known deleterious impact of social determinants on health outcomes, we expected to find differences in survey responses based on financial solvency, and we expected financial solvency reports to be consistent with traditional area income boundaries in our area (i.e., those in low income neighborhoods reporting the lowest rates of financial solvency). Finally, across surveys (and after the implementation of stay-at-home orders), we aimed to gather information about the extent to which individuals were seeking and receiving services and assistance with social needs impacted by the pandemic across the three main domains of SDOH, access to healthcare services and managing their own and their family health; impact of their neighborhood and built environment; and economic stability.

2 Related Work

2.1 What Are SDOH?

Over the last 30 years, researchers have focused on understanding the critical impacts that social factors have on our everyday health and wellness [35–39]. Social determinants of health are the "conditions in the environments where people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality of life outcomes and risks" [31]. The individual determinants include education access and quality, economic stability, healthcare access and quality, social connections, and neighborhoods/the built environment. Within each of these domains, there are a host of more granular issues at play (see Fig. 1).

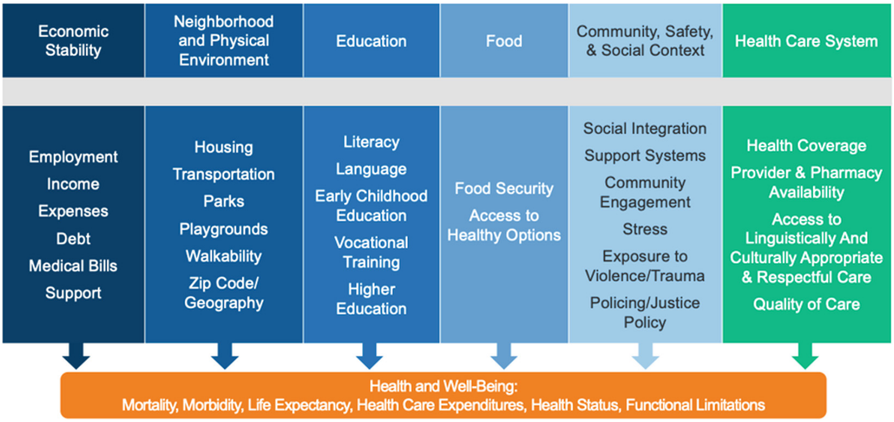


Fig. 1. Social Determinants of Health [33]

Research shows that resources that enhance the quality of life can have profound outcomes at the population level [32]. The World Health Organization estimates that SDOH account for between 30–55% of all health outcomes [34, 59].

2.2 SDOH and HCI

The pervasive health community and HCI community has conducted research that explores almost every social determinant. These include nutrition [21, 27, 28], telehealth to expand access to healthcare [22,23], access to maternal healthcare [23,24], low-income community needs [24,25], transportation [26], and access to affordable prescription medication [29,30]. Additionally, research has focused on health monitoring and support [41,42] as well as home health aides [40]. These bodies of research focus on understanding the interplay or impacts of technology in better articulating or understanding issues within these domains, not SDOH as a whole. Recent scholarship has looked at SDOH in a more holistic manner. Debopadhaya et al., recently looked at how temporal analysis of SDOH are associated specifically with COVID-19 mortality using county-level data from across the United States [43]. During a powerful keynote presentation, Grimes-Parker discussing health equity from a global perspective, highlighting the potential and pitfalls of intelligent interfaces in support them [44]. Additionally, focus has been given to appropriate approaches needed in research to achieve equity [45]. The research presented here takes a different approach - activating a diverse community stakeholder group via a socio-technical systems approach to identify real-time needs within a community during an evolving community health emergency.

2.3 Surveying of SDOH During Times of Crisis

Many surveys have been conducted with regards to measuring SDOH. Within the general population, these have focused on people with diabetes [50], children [49], cancer patients [47], people going through menopause [46], and even populations of immigrants [48]. However, all of these surveys have not focused on a hyper-local context - taking into consideration multiple communities to better understand the sub-population of interest. With respect to surveying SDOH needs during COVID, recent research has assessed pregnant women [54], nursing homes [53], ethnic groups [52], and children with mental health needs [51]. Several have looked at location-based needs [52,57,58]. However, none of this research was conducted within the contexts of a socio-technical system, nor do they discuss how findings were operationalized by community organizations or governments to make decisions on how to support communities during the pandemic. The only examples of SDOH and COVID surveys helping coordinate care or response to needs are focused within international contexts [55,56].

3 Method

Two cross-sectional surveys were developed and administered to understand the relation between COVID-19 impacts and components of social determinants of health in Northeastern Indiana. This research was approved by the Parkview Institutional Review Board on March 27, 2020. Survey A was developed by a multi-disciplinary group convened by the Health Commissioner in early March 2020. This group included representatives from the following agencies: Allen County Health department, Parkview Healthcare system, local government officials, the Allen County Commissioner's Office, business leader associations, local philanthropic and not-for-profit groups, and civic leaders. After results from Survey A were analyzed a disseminated, refinements were made by representatives from the health department, research scientists, government officials and local philanthropic groups to address the evolving situation resulting in Survey B.

3.1 Survey Objective

The first survey (Survey A) was conducted prior to the state mandated shelter-in-place and was focused on understanding anticipated community resources to sustain basic needs for individuals facing economic insecurity. The second survey (Survey B) was conducted during the shelter-in-place time period and had the additional aims of determining if community needs were being met and where individuals were finding resources as well as ascertaining the levels of employment and barriers to accessing unemployment benefits. Racial and gender data were not collected in Survey A to engender a sense of privacy for respondents.

3.2 Participants and Procedure

Data were from two brief online surveys discussed above (Survey A and B). Respondents to Survey B were able to complete the survey on behalf of others due to feedback from community stakeholders regarding concerns about getting an increased response rate for elderly, Spanish and Burmese speaking populations. Respondents were completely anonymous. The survey was available on the Allen County Health Department's COVID-19 website and the link was shared via community organizations, listservs, and social media. Additionally, the surveys were promoted through television via press conferences and the local public broadcasting station's weekly COVID-19 programming[1]. We utilized data from March 13 to June 22, 2020, in which a total of 6,031 surveys with an Indiana zip code were collected (4,992 from Survey A and 1,039 from Survey B).

3.3 Measures

Financial Solvency. Respondents indicated the length of time that they could sustain their household without additional income (1–2 weeks, 3–4 weeks, 5–6 weeks, 7+ weeks). The 5–6 weeks and 7+ weeks responses were combined to create a 3-level categorical variable.

Healthcare Related Characteristics. Respondents reported their health insurance status, whether they had an established healthcare provider (for self and dependents), members of household with chronic health condition (yes/no), type of prescriptions from a list of 7 general categories of medication with a free text option, ability to pay for medical treatments, and potential barriers to access and pay for medications during the pandemic.

---

[1] https://www.pbs.org/video/coronavirus-a-live-community-forum-april-3-2020-ke73wk/.

Service Preferences (Survey A ONLY). Types of organizations providing assistance with utilities, food/water, and healthcare were listed for respondents to select their preferred type of organization for receiving each type of assistance. Organizations included community, government, and faith-based groups. Service providers (Survey B ONLY). Respondents indicated which type of organizations they used to obtain assistance with utilities, food/water, and healthcare. Types of organizations included community, government, and faith-based groups.

Employment Status (Survey B ONLY). Respondents reported their current levels of employment (the same as pre-COVID, the same as pre-COVID but furloughed, the same as pre-COVID but taking a leave of absence, different job than pre-COVID, not currently employed, or N/A), how compensation compared to pre-COVID if employed (the same amount, more pay, less pay), and income source if a job was recently lost (severance, unemployment insurance, no income/unemployment denied, no income/do not qualify for unemployment, no income/have not applied for unemployment yet). Respondents reported whether their spouse or partner was employed and, if not, whether they received unemployment benefits

Needs (Survey B ONLY). Respondents indicated the extent to which the following needs were being met: stress/emotional support, transportation, utilities, rent/mortgage, accessing unemployment funds, accessing mental healthcare, accessing healthcare, obtaining food/water, obtaining medications, access to the internet, access to information/curriculum/government forms online. The response set was a 5-point Likert type (1 = none, 2 = some, 3 = half, 4 = most, or 5 = all). For those needs that respondents sought support, respondents specified sources of support from the following options: doctor/local healthcare provider, non-profit organization, family/friends, local school, church, local government agency, 2-1-1 (specialists available by phone to locate local resources and services), or Internet. Respondents were asked if the local stay of evictions was lifted, would they be able to pay current and/or back rent/mortgage (response categories: 1 = yes/have the funds to pay, 2 = yes/have some of the funds to pay, 3 = no/I do not have the funds to pay, 4 = I am not at risk of being evicted for non-payment).

3.4 Data Analysis

Descriptive summary statistics were calculated. The primary independent variable of interest was a 3-level categorical variable indicating number of weeks that households could sustain financially without additional pay (1–2 weeks, 3–4 weeks, 5 or more weeks). These three groups were compared using chi-square tests of independence or one-way ANOVA tests, as appropriate, and post hoc pairwise comparisons made between groups. Analyses were conducted using SAS software version 9.4 (Cary, NC).

4 Results

A total of 4992 Indiana residents completed Survey A. Data validity checks resulted in removing 7 respondents (6 respondents indicated their household included 30 or more members, 1 respondent indicated 125 individuals would need eldercare). An additional 3.8% (189/4992) of respondents did not complete the financial security question and were removed. The final sample included 4796 respondents. A total of 1101 residents of Northeast Indiana initiated Survey B; however, only 1039 responses were included in analyses due to missing data on the financial solvency item. Nearly all respondents completed Survey B on their own behalf (93.7%, 974/1039) and most were White (87.9%, 913/1039). In most cases, data presented refer to "survey results," generally; specification of Survey A or Survey B can be discerned from number of respondents and/or specific questions on the survey as listed in the methods section.

4.1 Economic Stability

As shown in Table 1, 43.9% (Survey A) and 21.3% (Survey B) of respondents indicated that their household could sustain without additional pay for 1–2 weeks, 24.9% (Survey A) and 19.8% (Survey B) for 3–4 weeks, and 31.3% (Survey A) and 58.9% (Survey B) for 5 weeks or longer. An inspection of respondent zip codes provided an indication of where need could be anticipated. Typically, non-profit organizations focus attention in high poverty zip codes. However, low financial solvency, as self-reported in this survey, was not limited to these high poverty zip codes. Another 10 zip codes each had 100 or more respondents indicating limited financial solvency; in total, respondents in these zip codes accounted for 66.8% (1405/2105) of the low financial solvency group. On average, households that could sustain for 5 weeks or more had the fewest number of household members, minor dependents, and dependents with special needs. Also, about 33% of respondents from households able to sustain for 1–2 weeks reported that they would need to find childcare; whereas less than 20% of respondents from households able to sustain for 5 weeks or more reported that they would need to find childcare. Also, about 65% of the low financial solvency group were concerned about finding or purchasing food.

Overall, the rate of unemployment was 17.1% (148/868) for respondents and 20.3% (111/546) for significant others (as reported by respondents). However, only 36 respondents reported receiving unemployment compensation, and 23 reported that their significant other received unemployment. The majority (81.7%, 709/868) of respondents had the same job as prior to the COVID-19 pandemic, but for these that were employed, 13.4% (119/891) of respondents and 17.9% (78/435) of significant others were receiving less pay than prior to the pandemic.

Health and Healthcare Characteristics. Table 2 highlights the insecurities related to healthcare (including medical characteristics and medications).

Table 1. Characteristics of sample by financial solvency group

Household Characteristics	How many weeks could you financially sustain your household if your workplace closed and received no additional pay?			Group Comparison
	1-2 weeks	3-4 weeks	5+ weeks	p-value
Number in household, avg(SD)				
	Survey A - 2082 *Survey B - 221*	*Survey A - 1173* *Survey B - 206*	*Survey A - 1487* *Survey B - 612*	
Survey A	3.14^{a} (1.8)	3.3^{b} (1.7)	3.0^{c} (1.6)	<0.001
Survey B	3.7^{a} (3.7)	3.1^{b} (1.6)	2.9^{b} (2.2)	<0.001
Number of dependent children requiring substitute childcare				
Survey A	*n=2082*	*n=1173*	*n=1487*	<0.001
0	67.1%	74.5%	82.9%	
1	11.6%	10.7%	6.3%	
2	13.5%	10.7%	7.3%	
3+	$7.9\%^{a}$	$4.2\%^{b}$	$3.6\%^{c}$	
Number of dependent children				
Survey B	*n=221*	*n=206*	*n=612*	
0	47.1%	50.0%	61.4%	0.004
1	20.4%	19.4%	14.1%	
2	21.3%	18.9%	13.7%	
3+	$11.3\%^{a}$	$11.7\%^{a}$	$29.9\%^{b}$	
Number of individuals with special needs				
Survey A	*n=2082*	*n=1173*	*n=1487*	
0	84.6%	88.8%	91.8%	0.01
1+	$15.4\%^{a}$	$11.2\%^{ab}$	$8.2\%^{b}$	
Survey B	n=221	n=206	n=612	
0	90.5%	90.3%	96.4%	<0.001
1+	$9.5\%^{a}$	$9.7\%^{a}$	$3.6\%^{b}$	
Worried about finding/purchasing food				
Survey A	*n=510*	*n=309*	*n=393*	
	$64.5\%^{a}$	$46.0\%^{b}$	$23.7\%^{c}$	<.001
Unemployed				
Survey B	*n=190*	*n=175*	*n=503*	
	16.3%	16.0%	17.7%	0.84
Significant and other unemployed				
Survey B	*n=102*	*n=111*	*n=333*	
	20.6%	17.01%	21.3%	0.63
Received stimulus check				
Survey B	*n=198*	*n=184*	*n=560*	
	$71.7\%^{a}$	$63.6\%^{ab}$	$57.1\%^{b}$	0.002
Risk of eviction/ability to pay current and back rent or mortgage				
Survey B	*n=163*	*n=161*	*n=498*	
0	42.9%	48.5%	56.6%	<0.001
1	14.1%	4.4%	0.6%	
2	14.1%	8.7%	1.0%	
3+	$28.8\%^{a}$	$38.5\%^{b}$	$41.8\%^{c}$	

Note: Superscripts indicate group differences such that those with different subscripts are significantly different ($p < 0.05$) from one another, and those with the same subscript are not significantly different.

Interestingly, about 90% of respondents that reported the lowest financial security had health insurance, which could have included Medicaid or Medicare, but only 40% (Survey A) and 43% (Survey B) of this group had the financial resources to pay for medical treatment. Leading up to the stay-at-home orders, 42% (Survey A) were worried about finding or purchasing medications and 6.3% (49/777) of the sample had issues obtaining medications since the beginning of the COVID-19 pandemic (Survey B), the rate of problems disproportionately affected respondents in the 1–2 week financial solvency group. The top barriers to obtaining medications were 'Not able to have an appointment with my physician or provider' and 'Can't pay for medications.' A few respondents reported seeking medical care in the Emergency Department (n = 12) and Urgent Care (n = 12) due to inability to obtain their medications.

The most common medications prescribed were for treating mental health concerns and regulating blood pressure across the three financial groups. In a test of independence, prescriptions for these two medications was not independent of financial solvency group. However, slightly different patterns emerged. The lowest financial solvency group reported highest rate of prescriptions for mental health concerns, and the highest financial solvency group reported highest rate of blood pressure medications. As shown in Table 2, across the three financial groups, the top medications prescribed were for treating mental health concerns and regulating blood pressure.

Sources of Assistance. Respondents were asked to report where they might seek assistance for key issues within three social determinants of health domains: health and healthcare; neighborhood and built environment; and economic stability. Across all respondents there was a preference towards accessing government resources for assistance with utilities or healthcare; however, they preferred community organizations for assistance with food and water. Respondents with less financial solvency (1–2 weeks) were more likely to report that they would rely on government resources for assistance with all needs when compared with those with greater reported financial security (see Table 3).

Respondents were asked where they had found assistance for key issues within three social determinants of health domains: health and healthcare; neighborhood and built environment; and economic stability. All groups primarily relied on family and friends for help. Respondents in the 5 or more weeks financial solvency group tended to rely on resources from church more than the 1–2 week financial solvency group. While those with less financial security (1–2 weeks solvency) reported they received more support from non-for-profit organizations (See Table 4).

Overall, respondents across all levels of financial solvency reported that at least half to most of their needs were being met. The highest category of needs was access to mental healthcare and stress/emotional support (see Table 5). As shown in Table 5, needs that were being met to a lesser extent were access to unemployment funds, and access of information and forms online. These gaps are most apparent among the 1–2 week financial solvency group.

Table 2. Health and Healthcare characteristics of sample by financial solvency group

Medical Characteristics	**How many weeks could you financially sustain your household if your workplace closed and received no additional pay?**			**Group Comparison**
	1-2 weeks	**3-4 weeks**	**5+ weeks**	**p-value**
Established healthcare provider for self				
	Survey A - 2103 *Survey B - 163*	*Survey A - 1191* *Survey B - 159*	*Survey A - 1496* *Survey B - 501*	
Survey A	78.7%^{a}	85.1%^{b}	87.9%^{c}	<0.001
Survey B	87.7%^{ab}	82.4%^{a}	90.4%^{b}	0.02
Established healthcare provider for dependents				
	Survey A - 2022 *Survey B - 164*	*Survey A - 1137* *Survey B - 160*	*Survey A - 1432* *Survey B - 502*	
Survey A	78.0%^{a}	82.0%^{b}	81.9%^{b}	0.004
Survey B	87.2%^{a}	87.3%^{a}	94.2%^{b}	0.02
Health insurance				
	Survey A - 2041 *Survey B - 117*	*Survey A - 1151* *Survey B - 110*	*Survey A - 1379* *Survey B - 294*	
Survey A	89.8%^{a}	93.5%^{b}	97.4%^{c}	<0.001
Survey B	89.6%^{a}	88.8%^{a}	95.6%^{b}	0.02
Financial resources to pay for medical treatment for self or household member				
	Survey A - 2016 *Survey B - 164*	*Survey A - 1135* *Survey B - 159*	*Survey A - 1429* *Survey B - 501*	
Survey A	40.6%^{a}	69.3%^{b}	88.9%^{c}	<0.001
Survey B	42.7%^{a}	59.1%^{b}	89.8%^{c}	<0.001
Worried about finding or purchasing medication for self or household member				
	n=475	*n=297*	*n=371*	
Survey A	41.7%^{a}	26.0%^{b}	16.7%^{c}	<.001
Have not been able to obtain medication				
	n=140	*n=148*	*n=489*	
Survey B	12.9%^{a}	8.8%^{a}	3.7%^{b}	<.001
Medication Type	**How many weeks could you financially sustain your household if your workplace closed and received no additional pay?**			**Group Comparison**
	1-2 weeks	**3-4 weeks**	**5+ weeks**	**p-value**
Survey B	*n=162*	*n=159*	*n=496*	
None	21.6%	22.6%	24.8%	0.66
Mental Health	51.9%^{a}	44.0%^{a}	31.3%^{b}	<0.001
Blood Pressure	40.7%	45.9%	37.3%	0.15
Inhalers	26.5%^{a}	23.9%^{a}	16.5%^{b}	0.008
Pain Medication	19.1%^{a}	10.7%^{b}	9.1%^{b}	0.002
Diabetes	17.3%	21.4%	15.5%	0.23
Blood Thinners	10.2%	17.1%	14.1%	0.32
Heart Failure	5.6%	4.4%	4.2%	0.78
Oxygen	4.3%^{a}	1.9%^{ab}	0.6%^{b}	0.004
Cancer Therapies	1.2%	0.0%	2.6%	0.08

Note: Superscripts indicate group differences such that those with different subscripts are significantly different ($p < 0.05$) from one another, and those with the same subscript are not significantly different.

Table 3. Counts of preferences for obtaining resources for utilities, food/water, and healthcare from government, community, and faith-based organizations (Survey A Only)

What local resources would you use if you needed help with the items below?	**How many weeks could you financially sustain your household if your workplace closed and received no additional pay?**			**Group Comparison**
	1-2 weeks	**3-4 weeks**	**5+ weeks**	**p-value**
Utilities				
	n=1933	*n=1060*	*n=1323*	
Government	26.9%^{a}	23.3%^{b}	19.1%^{c}	<0.001
Community Orgs	15.0%	14.3%	12.3%	0.10
Faith-Based	14.0%^{a}	17.9%^{b}	16.2%^{b}	0.01
Food/Water				
	n=1958	*n=1102*	*n=1373*	
Government	21.7%^{a}	17.9%^{b}	16.2%^{b}	<0.001
Community Orgs	26.5%^{a}	30.6%^{b}	29.6%^{b}	0.03
Faith-Based	23.3%	26.0%	25.2%	0.21
Healthcare				
	n=1971	*n=1114*	*n=1380*	
Government	28.5%^{a}	26.8%^{a}	21.9%^{b}	<0.001
Community Orgs	9.2%	7.6%	8.0%	0.25
Faith-Based	6.9%	7.5%	5.9%	0.31

Note: Superscripts indicate group differences such that those with different subscripts are significantly different ($p < 0.05$) from one another, and those with the same subscript are not significantly different.

Table 4. Resources used to address household needs (utilities, healthcare, food, transportation, emotional support, rent/mortgage, unemployment funds, internet, information) by financial solvency group (Survey B Only).

For the needs you have listed, have you received support from the following:	**How many weeks could you financially sustain your household if your workplace closed and received no additional pay?**			**Group Comparison**
	1-2 weeks	**3-4 weeks**	**5+ weeks**	**p-value**
	n=122	*n=113*	*n=345*	
Doctor or local healthcare provider	53.3%	46.9%	47.5%	0.51
Non-profit organization	18.0%^{a}	9.7%^{ab}	4.9%^{b}	<0.001
Family and friends	67.2%	75.2%	65.8%	0.17
Local school(s)	6.6%	15.0%	8.7%	0.06
Church	6.6%^{a}	11.5%^{a}	22.0%^{b}	<0.001
Local government agency	3.3%	1.8%	2.0%	0.68
2-1-1 (state-based resource hotline)	1.6%^{ab}	2.7%^{a}	0.0%^{b}	0.007
Internet resources	38.5%^{a}	36.2%^{a}	49.6%^{b}	0.01

Note: Superscripts indicate group differences such that those with different subscripts are significantly different ($p < 0.05$) from one another, and those with the same subscript are not significantly different.

Table 5. Mean level of household needs (utilities, healthcare, food, transportation, emotional support, rent/mortgage, unemployment funds, internet, information) by financial solvency group (Survey B Only).

Mean level of household needs that were met	How many weeks could you financially sustain your household if your workplace closed and you received no additional pay? 1-2 weeks		3-4 weeks		5+ weeks		Group Comparison p-value
	n	*Mean*	*n*	*Mean*	*n*	*Mean*	
Economic Stability							
Access to unemployment funds	23	3.83	20	4.30	47	4.38	0.15
Obtaining food/water	156	4.40^a	155	4.65^b	466	4.85^c	<0.001
Utilities	151	4.43^a	155	4.76^b	459	4.92^c	<0.001
Rent/mortgage	133	4.43^a	147	4.77^b	387	4.95^c	<0.001
Transportation	144	4.72^a	144	4.85^a	447	4.94^a	<0.001
Health and Healthcare							
Access to mental healthcare	88	3.65^a	84	3.92^a	210	4.49^b	<0.001
Access to healthcare	130	4.02^a	130	3.98^a	396	4.40^b	<0.001
Obtaining medications	137	4.48^a	138	4.72^b	412	4.89^c	<0.001
Neighborhood and Built Environment							
Stress/emotional support	138	3.46^a	143	3.71^b	443	4.14^c	<0.001
Access to online support	43	3.53	45	3.53	75	3.60	0.85
Access to the internet	158	4.65^a	152	4.82^b	468	4.86^b	<0.001

Respondents used a 5-point Likert scale to indicate extent to which each need is being met, 1=NONE of my need is being met, 5=ALL of my need is being met. Superscripts indicate group differences such that those with different subscripts are significantly different ($p < .05$) from one another, and those with the same subscript are not significantly different.

Respondents reported negative effects of social distance for self and other household members, regardless of financial solvency group. However, only 15.2% (93/612) of entire sample sought mental health resources (see Table 6).

Table 6. Impact of social distancing on household members and mental health help by financial solvency group (Survey B Only).

Currently, is the social distancing and isolation having a negative effect on the mental health of the following:	**How many weeks could you financially sustain your household if your workplace closed and received no additional pay?**			**Group Comparison**
	1-2 weeks	**3-4 weeks**	**5+ weeks**	**p-value**
	n=133	*n=135*	*n=317*	
Myself	$72.9\%^a$	$68.8\%^b$	$60.3\%^b$	0.02
My spouse/partner	43.6%	446.4%	40.7%	0.53
My children	46.6%	44.0%	39.8%	0.37
Sought help for mental health impact	$21.2\%^a$	$10.8\%^b$	$14.5\%^b$	0.05

Note: Superscripts indicate group differences such that those with different subscripts are significantly different ($p < 0.05$) from one another, and those with the same subscript are not significantly different.

4.2 Overall Community Impacts

The results from the surveys were analyzed immediately and distributed back to the network of stakeholders that helped build the original survey. This lead to real-time, data-driven responses across the community at all levels. Figure 2 highlights some of these outcomes and the different stakeholder groups that were impacted by the data[2].

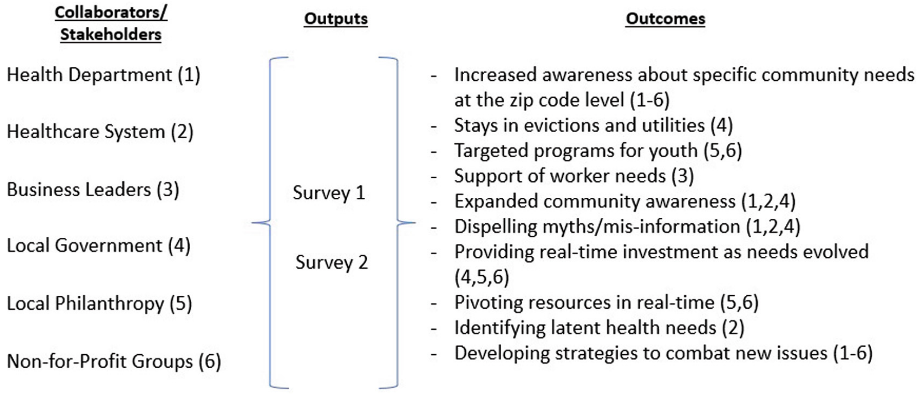

Collaborators/ Stakeholders
Outputs
Outcomes
Health Department (1)
Healthcare System (2)
Business Leaders (3)
Local Government (4)
Local Philanthropy (5)
Non-for-Profit Groups (6)
Survey 1
Survey 2
- Increased awareness about specific community needs at the zip code level (1-6)
- Stays in evictions and utilities (4)
- Targeted programs for youth (5,6)
- Support of worker needs (3)
- Expanded community awareness (1,2,4)
- Dispelling myths/mis-information (1,2,4)
- Providing real-time investment as needs evolved (4,5,6)
- Pivoting resources in real-time (5,6)
- Identifying latent health needs (2)
- Developing strategies to combat new issues (1-6)

Fig. 2. Impact of Survey Findings on Different Stakeholder Groups

5 Discussion

In response to emerging needs within our community due to the COVID-19 pandemic, leaders from the areas of public health, health systems, government entities, local business, and philanthropy convened and developed a sequence of two community surveys. The community-based surveys revealed consistent trends regarding the widespread impact of COVID-19 on social determinants of health and also provided the impetus for collaborative community efforts towards resource allocation. The results were used by public health officials, city government, and other community organizations to develop wide-scale, targeted efforts in response to the pandemic that affected not only disease spread and mortality but also other aspects of living, such as the ability to travel, work outside the home, and receive standard and emergency medical and psychological treatment.

These findings are useful from a research perspective because they identify SDOH disparities through the lens of financial solvency and provide a model for informatics professionals to lend their skill set to community agencies during a crisis. In line with previous studies of SDOH [14–17], respondents who were less financially solvent reported inequalities in almost every area (e.g., unemployment, children in home requiring care, medications taken, and healthcare and insurance characteristics) than those with more financial solvency. Those in the lowest financial solvency group also reported more SDOH-related needs and willingness to engage with community organizations helping with economic stability, health and healthcare, and neighborhood and built environment.

---

[2] https://www.inputfortwayne.com/features/community-needs-mirro.aspx.

Interestingly, and in contrast to our prediction, some of these financially insecure families were outside of traditional low-income areas. Thus, mitigation and prevention efforts using geocoding to help address pandemic- or other crisis-related needs (e.g., [7,9]) should not rely simply on county residency or zip codes for their targeted efforts. In times when crises restrict movement, work, and/or childcare, SDOH needs transcend these traditional income boundaries. Activation and advertisement of community resources might be especially important for those facing financial solvency issues for the first time. Perhaps, these families may not have been familiar with the organizations or processes that would allow them to address some of their SDOH needs. This study provides a direction for future research and points to the importance of using community-based, data-driven approaches, preferably just-in-time analyses, and not simply relying on categorical data, like zip codes, when developing social policy and relief measures.

In an effort to measure a wide range of health needs, we also surveyed participants on their mental health issues. Many, especially those in the low financial solvency group, reported that the shelter-in-place mandates had a negative effect on their mental health and the mental health of their family members [60]. This finding reinforces assertions that mental health supports are critical during times of disaster or infectious disease outbreak [17]. Respondents also reported that less than half of their mental health needs were being met. Additionally, despite negative mental health effects of social distancing reported by many respondents, only 15% sought mental health resources during sheltering-in-place. Medications for mental health needs were also amongst the highest concern for those in the lowest (1–2 weeks) financial solvency group and the second highest for everyone else. This points to the importance of public policy that addresses the mental health needs along with the physical health needs of our population in times of pandemic or other public health crises.

Importantly, the just-in-time analyses of these data led to actionable plans for community resource allocation. This demonstrates the feasibility of community, grant-based collaborations, like those which have been used in non-pandemic times to address the needs of those who are at greatest disadvantage from a SDOH perspective [13]. Specifically, using the results from the first survey (Survey A), the local city government made legislative mandates to aid those with low financial solvency, including enacting an order to prevent utilities services discontinuance due to non-payment and a moratorium on evictions due to non-payment of rent. Philanthropic groups also used the results from Survey A to justify special grant awards prioritizing community organizations demonstrating strategies impacting identified areas of need.

There is evidence that these community need surveys, just-in-time analyses, and related coordinated action plans made an immediate impact [18]. From an individual perspective, prior to the shelter-in-place orders, respondents less financially secure anticipated the primary source of support coming from the government. However, after shelter-in-place took effect, this group reported community organizations being a primary source of support, which may reflect the actions taken by local community and philanthropic groups in response to our analyses of Survey A, including directing resources to help families pay for food, housing, and medical care. Meanwhile, from an organizational/municipal perspective, our initiative demonstrated that by quantifying constituent concerns, regional and local governments may be able to leverage that data for state and national funding. As an example, in our district, the Mayors and Commissioners Caucus of Northeast Indiana were able to develop a request for the Governor's Office that illustrated a \$300+M need for state assistance.

A limitation of this work includes potential under-representation of communities that are in the most need due to COVID-19. Because of the pandemic, traditional methods of access/canvasing were not possible when recruiting for the survey. Additionally, the methods do not allow us to parse the levels of needs of social determinants of health prior to the pandemic. However, as noted in the text, survey results uncovered zip codes where public data would indicate financial stability where the surveys uncovered lack of stability/fragility.

6 Conclusion

To effectively govern, it is important that government leaders have an ability to understand the current tenor of the community. During a pandemic, the ability to quantify constituent need is even more essential. The identification of those needs, via our community needs assessment surveys, allowed government officials to come together and conduct weekly press briefings for the public to work toward allaying some of the more predominant concerns. Additionally, in gaining a better understanding of the limits being placed on the lives of constituents, priorities were developed by city and county governments to allow for a smoother transition in the lives of those community members affected by the pandemic. Considering that COVID-19 disproportionately impacts various populations [19,20], our social determinants of health lens helped to disentangle the complexity of the impacts of this disease, which helped community leaders develop targeted interventions for those most in need.

References

1. Leigh, J.P., et al.: A national cross-sectional survey of public perceptions, knowledge, and behaviors during the COVID-19 pandemic. PLoS ONE **15**(10), e0241259 (2020)
2. Jung, S., Kneer, J., Drueger, T.: The German COVID-19 survey on mental health: primary results. medRxiv (2020). https://doi.org/10.1101/2020.05.06.20090340
3. Hubner, C., Bruscatto, M., Lima, R.: Distress among Brazilian university students due to the Covid-19 pandemic: survey results and reflections. medRxiv (2020). https://doi.org/10.1101/2020.06.19.20135251
4. Hooper, M.W., Nápoles, A.M., Pérez-Stable, E.J.: COVID-19 and racial/ethnic disparities. JAMA **323**(24), 2466–2467 (2020). https://doi.org/10.1001/jama.2020.8598

5. Kakol, M., Upson, D., Sood, A.: Susceptibility of southwestern American Indian tribes to coronavirus disease 2019 (COVID-19). J. Rural Health **37**(1), 197–199 (2020). https://doi.org/10.1111/jrh.12451
6. Khalatbari-Soltani, S., Cumming, R.G., Delpierre, C., Kelly-Irving, M.: Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards. J. Epidemiol. Community Health **74**(8), 620–623 (2020). https://doi.org/10.1136/jech-2020-214297
7. Chin, T., et al.: US county-level characteristics to inform equitable COVID-19 response. medRxiv (2020). https://doi.org/10.1101/2020.04.08.20058248
8. Wadhera, R.K., et al.: Variation in 300 COVID-19 hospitalizations and deaths across New York City boroughs. JAMA **323**(21), 2192–2195 (2020). https://doi.org/10.1001/jama.2020.7197
9. Krieger, N., Waterman, P.D., Chen, J.T.: COVID-19 and overall mortality inequities in the surge in death rates by zip code characteristics: Massachusetts, January 1 to May 19, 2020. Am. J. Public Health **110**(12), 1850–1852 (2020). https://doi.org/10.2105/AJPH.2020.305913
10. Mahajan, U.V., Larkins-Pettigrew, M.: Racial demographics and COVID-19 confirmed cases and deaths: a correlational analysis of 2886 US counties. J. Public Health **42**(3), 445–447 (2020). https://doi.org/10.1093/pubmed/fdaa070
11. Kayman, H., Ablorh-Odjidja, A.: Revisiting public health preparedness: incorporating social justice principles into pandemic preparedness planning for influenza. J. Public Health Manag. Pract. **12**(4), 373–380 (2006)
12. Bonaccorsi, G., et al.: Economic and social consequences of human mobility restrictions under COVID-19. Proc. Natl. Acad. Sci. U.S.A. **117**(27), 15530–15535 (2020). https://doi.org/10.1073/pnas.2007658117
13. Baril, N., Patterson, M., Boen, C., Gowler, R., Norman, N.: Building a regional health equity movement: the grantmaking model of a local health department. Family Community Health J. Health Promot. Maintenance **34**(Suppl 1), S23–S43 (2011). https://doi.org/10.1097/FCH.0b013e318202a7b0
14. Bucciardini, R., et al.: The health equity in all policies (HEiAP) approach before and beyond the Covid-19 pandemic in the Italian context. Int. J. Equity Health. **19**(1), 1–3 (2020). https://doi.org/10.1186/s12939-020-01209-0
15. Rudner, N.: Disaster care and socioeconomic vulnerability in Puerto Rico. J. Health Care Poor Underserved **30**(2), 495–501 (2019). https://doi.org/10.1353/hpu.2019.0043
16. Thornton, R.L., Glover, C.M., Cené, C.W., Glik, D.C., Henderson, J.A., Williams, D.R.: Evaluating strategies for reducing health disparities by addressing the social determinants of health. Health Aff. **35**(8), 1416–1423 (2016). https://doi.org/10.1377/hlthaff.2015.1357
17. McFarlane, A.C., Williams, R.: Mental health services required after disasters: learning from the lasting effects of disasters. Depress. Res. Treat. (2012). https://doi.org/10.1155/2012/970194
18. Elser, H., et al.: Implications of the COVID-19 San Francisco Bay area shelter-in-place announcement: a cross-sectional social media survey. medRxiv (2020). https://doi.org/10.1101/2020.06.29.20143156
19. Guo, W., et al.: Quick community survey on the impact of COVID-19 outbreak for the healthcare of people living with HIV. Zhonghua Liu Xing Bing Xue Za Zhi **41**(5), 663–667 (2020). https://doi.org/10.3760/cma.j.cn112338-20200314-00345
20. Alon, T.M., Doepke, M., Olmstead-Rumsey, J., Tertilt, M.: The impact of COVID-19 on gender equality (no. w26947). National Bureau of Economic Research (2020). https://www.nber.org/papers/w26947

21. Ylizaliturri-Salcedo, M.Á., García-Macías, J.A., Aguilar-Noriega, L., Cárdenas-Osuna, R.: What did you eat today? Designing a health program on nutritional poverty. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2018), pp. 274–279. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3240925.3240945
22. Pater, J.A., Coupe, A., Miller, A.D., Reining, L.E., Drouin, M., Toscos, T.: Design opportunities and challenges for app-based telemental health technologies for teens and young adults. In: Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2020), pp. 188–200. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3421937.3422016
23. Chaudhry, B.M., Faust, L., Chawla, N.V.: Towards an integrated mHealth platform for community-based maternity health workers in low-income communities. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2018), pp. 118–127. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3240925.3240938
24. Al Mahmud, A., Keyson, D.V.: Designing with midwives: improving prenatal care in low resource regions. In: Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2013). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, pp. 180–183 (2013). https://doi.org/10.4108/icst.pervasivehealth.2013.252032
25. Barnes, P., Caine, K., Connelly, K., Siek, K.: Understanding the needs of low SES patients with type 2 diabetes. In: Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2013). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, pp. 302–306 (2013). https://doi.org/10.4108/icst.pervasivehealth.2013.252153
26. Dillahunt, T.R., Veinot, T.C.: Getting there: barriers and facilitators to transportation access in underserved communities. ACM Trans. Comput.-Hum. Interact. **25**(5), 39 (2018). Article 29. https://doi.org/10.1145/3233985
27. Dillahunt, T.R., Simioni, S., Xu, X.: Online grocery delivery services: an opportunity to address food disparities in transportation-scarce areas. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI 2019), pp. 1–15. Association for Computing Machinery, New York (2019). Paper 649. https://doi.org/10.1145/3290605.3300879
28. Grimes, A., Bednar, M., Bolter, J.D., Grinter, R.E.: EatWell: sharing nutrition-related memories in a low-income community. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW 2008), pp. 87–96. Association for Computing Machinery, New York (2008). https://doi.org/10.1145/1460563.1460579
29. Osmani, V., Forti, S., Mayora, O., Conforti, D.: Enabling prescription-based health apps. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2017), pp. 272–275. Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3154862.3154911
30. Livet, M., Levitt, J.M., Lee, A., Easter, J.: The pharmacist as a public health resource: expanding telepharmacy services to address social determinants of health

during the COVID-19 pandemic. Exploratory Res. Clin. Soc. Pharm. **2**, 100032 (2021)

31. ODPHP: Healthy People 2030: Building a healthier future for all. Office of Disease Prevention and Health Promotion, U.S. Department of Health and Human Services. https://health.gov/healthypeople
32. CDC-SDOH - US Centers for Disease Control. https://www.cdc.gov/socialdeterminants/index.htm. Accessed 13 Aug 2022
33. Drake, P., Rudowitz, R.: Tracking Social Determinants of Health During the COVID-19 Pandemic. The Kaiser Family Foundation, 21 April 2022. https://www.kff.org/coronavirus-covid-19/issue-brief/tracking-social-determinants-of-health-during-the-covid-19-pandemic/
34. Social Determinants of Health - World Health Organization. https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1. Accessed 13 Aug 2022
35. Marmot, M., Bell, R.: Fair society, healthy lives. Public Health **126**(Suppl 1), S4–S10 (2012)
36. Braveman, P., Egerter, S., Williams, D.R.: The social determinants of health: coming of age. Annu. Rev. Public Health **32**, 381–98 (2011)
37. Kaplan, G.A., Shema, S.J., Leite, C.M.: Socioeconomic determinants of psychological well-being: the role of income, income change, and income sources during the course of 29 years. Ann. Epidemiol. **18**, 531–7 (2008)
38. Lynch, J.W., Everson, S.A., Kaplan, G.A., Salonen, R., Salonen, J.T.: Does low socioeconomic status potentiate the effects of heightened cardiovascular responses to stress on the progression of carotid atherosclerosis? Am. J. Public Health **88**, 389–94 (1998)
39. Adler, N.E., Marmot, M., McEwen, B.S., Stewart, J. (eds.): Socioeconomic Status and Health in Industrial Nations: Social, Psychological, and Biological Pathways. New York Academy of Sciences, New York (1999)
40. Al-Masslawi, D., et al.: SuperNurse: nurses' workarounds informing the design of interactive technologies for home wound care. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2017), pp. 193–202. Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3154862.3154865
41. Gupta, A., Heng, T., Shaw, C., Li, L., Feehan, L.: Towards developing an e-coach to support arthritis patients in maintaining a physically active lifestyle. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2018), pp. 392–395. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3240925.3240954
42. Parsons, A., Chung, C.-F., Donohue, M., Munson, S.A., Seibel, E.J.: Opportunities for oral health monitoring technologies beyond the dental clinic. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2018), pp. 327–335. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3240925.3240973
43. Debopadhaya, S., Erickson, J.S., Bennett, K.P.: Temporal analysis of social determinants associated with COVID-19 mortality. In: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB 2021), pp. 1–10. Association for Computing Machinery, New York (2021). Article 41. https://doi.org/10.1145/3459930.3469535
44. Parker, A.G.: Achieving health equity: the power & pitfalls of intelligent interfaces. In: 26th International Conference on Intelligent User Interfaces (IUI 2021), pp. 5–6. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3397481.3457413

45. Sabharwal, A., Barua, S., Kerr, D.: A systems approach to achieve equity in healthcare research. GetMobile Mob. Comput. Commun. **25**(3), 5–11 (2022). https://doi.org/10.1145/3511285.3511287
46. Namazi, M., Sadeghi, R., Behboodi Moghadam, Z.: Social determinants of health in menopause: an integrative review. Int. J. Women's Health. **11**, 637–647 (2019). PMID: 31849539. PMCID: PMC6910086. https://doi.org/10.2147/IJWH.S228594
47. Kurani, S.S., McCoy, R.G., Lampman, M.A., et al.: Association of neighborhood measures of social determinants of health with breast, cervical, and colorectal cancer screening rates in the US Midwest. JAMA Netw. Open **3**(3), e200618 (2020). https://doi.org/10.1001/jamanetworkopen.2020.0618
48. Chang, C.D.: Social determinants of health and health disparities among immigrants and their children. Curr. Probl. Pediatr. Adolesc. Health Care **49**(1), 23–30 (2019)
49. Sokol, R., et al.: Screening children for social determinants of health: a systematic review. Pediatrics **144**(4), e20191622 (2019)
50. Hill-Briggs, F., et al.: Social determinants of health and diabetes: a scientific review. Diabetes Care **44**(1), 258–279 (2021)
51. Xiao, Y., Yip, P.S., Pathak, J., Mann, J.J.: Association of social determinants of health and vaccinations with child mental health during the COVID-19 pandemic in the US. JAMA Psychiat. **79**(6), 610–621 (2022). https://doi.org/10.1001/jamapsychiatry.2022.0818
52. Bauer, C., et al.: Census tract patterns and contextual social determinants of health associated with COVID-19 in a hispanic population from South Texas: a spatiotemporal perspective. JMIR Public Health Surveill. **7**(8), e29205 (2021). https://doi.org/10.2196/29205
53. Hege, A., Lane, S., Spaulding, T., Sugg, M., Iyer, L.S.: County-level social determinants of health and COVID-19 in nursing homes, United States, June 1, 2020–January 31, 2021. Public Health Rep. **137**(1), 137–148 (2022)
54. Prasannan, L., et al.: Social determinants of health and coronavirus disease 2019 in pregnancy. Am. J. Obstet. Gynecol. MFM **3**(4), 100349 (2021)
55. Ataguba, O.A., Ataguba, J.E.: Social determinants of health: the role of effective communication in the COVID-19 pandemic in developing countries. Glob. Health Action **13**(1), 1788263 (2020). https://doi.org/10.1080/16549716.2020.1788263
56. Sharma, S., Walton, M., Manning, S.: Social determinants of health influencing the New Zealand COVID-19 response and recovery: a scoping review and causal loop diagram. Systems **9**(3), 52 (2021)
57. Moise, I.K.: Peer reviewed: variation in risk of COVID-19 infection and predictors of social determinants of health in Miami-Dade County, Florida. Prev. Chronic Dis. **17**, E124 (2020)
58. Baker, D.R., Cadet, K., Mani, S.: COVID-19 testing and social determinants of health among disadvantaged Baltimore neighborhoods: a community mobile health clinic outreach model. Popul. Health Manag. **24**(6), 657–663 (2021)
59. Magnan, S.: Social Determinants of Health 101 for Health Care: Five Plus Five. NAM Perspectives. Discussion Paper, National Academy of Medicine, Washington, DC (2017)
60. Drouin, M., McDaniel, B.T., Pater, J., Toscos, T.: How parents and their children used social media and technology at the beginning of the COVID-19 pandemic and associations with anxiety. Cyberpsychol. Behav. Soc. Netw. **23**(11), 727–736 (2020). https://doi.org/10.1089/cyber.2020.0284

Machine Learning, Human Activity Recognition and Speech Recognition

Up-Sampling Active Learning: An Activity Recognition Method for Parkinson's Disease Patients

Peng Yue[1], Xiang Wang[2], Yu Yang[2], Jun Qi[3], and Po Yang[1](✉)

[1] Department of Computer Science Faculty of Engineering, University of Sheffield, Sheffield, UK
Po.Yang@sheffield.ac.uk
[2] National Pilot School of Software, Yunnan University, Kunming, China
[3] Department of Computing, Xi'an JiaoTong-Liverpool University, Suzhou, China
Jun.Qi@xjtlu.edu.cn

Abstract. Parkinson's Disease (PD) is the second most common neurodegenerative disease. With the advancement of technologies of big data, wearable sensing and artificial intelligence, automatically recognizing PD patients' Physical Activities (PAs), health status and disease progress have become possible. Nevertheless, the PA measures are still facing challenges especially in uncontrolled environments. First, it is difficult for the model to recognize the PA of new PD patients. This is because different PD patients have different symptoms, diseased locations and severity that may cause significant differences in their activities. Second, collecting PA data of new PD patients is time-consuming and laborious, which will inevitably result in only a small amount of data of new patients being available. In this paper, we propose a novel up-sampling active learning (UAL) method, which can reduce the cost of annotation without reducing the accuracy of the model. We evaluated the performance of this method on the 18 PD patient activities data set collected from the local hospital. The experimental results demonstrate that this method can converges to better accuracy using a few labeled samples, and achieve the accuracy from 44.3% to 99.0% after annotating 25% of the samples. It provides the possibility to monitor the condition of PD patients in uncontrolled environments.

Keywords: Activity Recognition · Active Learning · Parkinson's Disease · Cross-Subject

1 Introduction

Parkinson's Disease (PD), one of the common chronic diseases in the elderly, is a progressive neurological disorder caused by the loss of dopamine-producing nerve cells in the brain [1]. The estimated number of PD patients will reach 8 million in 2030 [2]. One of the main challenges is how to monitor the condition of PD patients for a long time to improve efficiently the healthcare of these patients. In the course of treatment,

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 229–246, 2023.
https://doi.org/10.1007/978-3-031-34586-9_16

the dosage of the medicine given by the doctor usually vary with the patient's condition. Especially the motor state of the PD patients in late stage may generate different symptoms from patient to patient [3–5]. Usually, patients need to go to the hospital to assess their condition under the guidance of neurologists. In view of the high correlation between the human Physical Activity (PA) index and the PD diagnosis score, the Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) is utilised by physicians for the diagnosis of functional symptoms of PD in the early attempts, such as tremor and postural balance disorders [6]. Physical activity recognition (PAR) can provide clinicians with quantitative profiles of motor function behavior in natural environments in long-term period, which can further make treatment strategies objectively adapted.

However, due to the complicated and subjective scoring process, clinically automatic identification of the PAR for PD patients is urgently needed to reduce requirements for physicians as well as save time and effort. As such, researches on objective scoring via computer-aided diagnosis systems have been laid considerably theoretical foundations in previous studies [7–11]. Among them, most attentions focus on either designing standalone novel wearable sensors to achieve highly accurate PAR [12–14], or investigating advanced machine learning algorithms for training features from observed wearable sensory data of human body positions into specific several PA subjects [15–17]. Additionally, some researchers investigate how to attach wearable sensors into the most suitable positions for the best performance [18–21]. While these typical technologies have been capable of achieving satisfactory results, the majority concern on controlled environments (e.g., hospitals) but not uncontrolled environments (e.g., home). However, most patients measure PA after medication in controlled environments and thus the results may impact the overall diagnosis of the symptom [22]. Therefore, lifelong PA monitoring in an uncontrolled environment is extremely significant.

Nevertheless, measuring lifelogging PA currently may have the following obstacles that would affect their practical usefulness. First, in order to achieve high performance, some specific wearable sensors have to be distributed over the human body [23–25], but the accuracy would usually be greatly reduced by using consumer wearable devices in uncontrolled environments. Second, the types of PA are predefined and verified in manual-designed controlled environments, but for the PA data in the uncontrolled environment, PD patients may not be required to accurately record and label data samples for such a long time [26, 27]. Third, the individual conditions and life patterns of each patient are different, and thus the accuracy of PA measurement would be greatly affected by these uncertain factors, making traditional methods inapplicable. Thus, PAR for PD diagnosis in uncontrolled environments is still a very challenging issue [28].

Targeting on the challenges above, we propose a novel up-sampling active learning (UAL) framework by using only 5%–20% of the labeled samples to achieve the training effect of using all the labeled data. The ensemble classifier is exploited to train PA data of the PD patients collected from our local hospital. Under the circumstances that minimize the annotation cost, we explored the effect of different sampling strategies and up-sampling on active learning performance that is able to dynamically discover new patient PA. The proposed method is capable of adapting PD patients' PAR and compared the effects of different sampling strategies and up-sampling on the experiment results.

The experimental results demonstrate that by using the active learning of Best-versus-Second Best (BvSB) sampling strategy, the model can converge to the better accuracy. For the 6 new PD patients in the test set, after annotating 5.63% to 25.00% of the data, the model accuracy can reach 99%. For the more severe PD patients, the proportion of labeled samples required to achieve 99.0% classification accuracy is 17.26% with 23.15% samples without annotation.

The remainder of this paper is organized as follows. Section 2 reviews related work on activity recognition. We describe the details of data acquisition in Sect. 3, the Sect. 4 introduces the details of our proposed method up-sampling active learning (UAL), the experimental results are in the Sect. 5, and the Sect. 6 states our conclusions and future work.

2 Related Work

PAR leveraging advanced sensors has been investigated extensively over the past decade. Among the sensor-based methods, devices such as accelerometers, gyroscopes, magnetometers, ambient sensors, and RFID tags are widely used [29–31]. These technologies have shown great potential and are beneficial to medical care services such as clinical interventions for the elderly and patients with chronic diseases [7, 9, 32, 33]. In recent years, increasing number of researches have focused on PAR for PD diagnosis in naturalistic settings with easy-to-easy devices such as mobile phones or wrist band. Cheng et al. [34] proposed a deep neural network to distinguish gait activities (walking, jogging, etc.) from stationary activities (sitting, standing, etc.) between PD patients and healthy subjects with 98% of accuracy. Albert et al. [35] used standard machine learning algorithms to recognize PAs of PD patients and healthy subjects respectively and analysed the different features of PA between PD patients and healthy subjects. MPower [36], a typical mobile app developed using Apple’s ResearchKit library, is to remotely monitor PAs of PD patients and then analyze daily changes in their symptom severity and medicine conditions. The platform is a promise to conduct in naturalistic environments.

Deep learning is robust to noise and has high classification accuracy compared with traditional machine learning methods. The study [37] combined Linear Discriminant Analysis (LDA) and Long Short Term Memory (LSTM), and performed non-steady-state circuit tests including stairs, slopes and direction changes in mild PD patients and healthy subjects, achieved 80% F1 score when using LSTM only for the lower body data. However, deep learning requires a large amount of data, large amount of calculation, high cost and long time. Research [38] solves the problem that CNN is not suitable for small-scale and large-scale intra-class noise data sets through data enhancement. Using the enhanced data for model training, the recognition accuracy of CNN on 25 PD patient data reached 86.88%, and the accuracy of CNN was 7% higher than standard machine learning methods.

On the other hand, supervised learning methods require a large number of correct annotated data sets for training, which is often difficult to achieve in a naturalistic environment. Conventional inertial sensor data is difficult to annotate, so regardless of the actual experimental environment or the real environment, researchers generally use video to assist them in annotating [13]. In addition, self-recall and experience sampling are also commonly used to obtain labeled data, but such methods are prone to errors. In order to reduce the amount of data required by the classifier, semi-supervised learning [39, 40] and transfer learning [41–43] have also been widely used in the field of PAR. The semi-supervised method predicts unlabeled samples, selects the samples that the classifier has the most confidence in the prediction results, and then adds them to the training set for training. One drawback to this is that the accuracy of the classifier will be reduced when incorrect predictions are added to the training set. Transfer learning leverages the knowledge learned from source domain to predict the target domain which has a small number of labels. This method usually requires that the source domain and the target domain are very similar, which is often difficult to achieve in the real setting.

In contrast to semi-supervised learning, active learning reduces the amount of data required by the model by detecting the most informative samples and asking users to query a label. This method improves the performance of the model and reduces the requirement of the model on the amount of data. A large body of work has used active learning to tackle the issue of label acquisition in activity recognition. For example, Stikic et al. explored and analyzed the effects of semi-supervised learning and active learning on reducing the label requirement of model [44]. Liu et al. analyzed the feasibility of active learning to search for the most informative samples to be labeled in activity recognition, by using the lowest confident level and high disagreement between two classifiers as the active sampling [45]. Study [46] derives a hierarchical Bayesian model, which combines active learning and transfer learning. Finally, they conclude that this method can use fewer tags on the target domain to achieve faster learning. AALO [47] label overlapped activities by active learning. The combination of deep learning and active learning has also been widely used to identify human activities [48–51]. However, there are few studies on the activity recognition of PD patients, especially for new PD patients. In this paper, for ensuring that the classifier converges to good accuracy with fewer annotated examples, we propose an up-sampling active learning method, which makes it possible to recognize the activity of PD patients in a real environment.

3 Data Collection

The participants were recruited from patients with PD symptoms who visited the First People's Hospital of Yunnan Province neurology clinic from August 2020 to January 2021. A total of 18 patients voluntarily participated in the experiment under the premise of obtaining the written informed consent of each patient. Neurologists label the PD patients with any of the (0–4) MDS-UPDRS score based on the intensity and prevalence of these motor symptoms. After being rated by a professional doctor, among the 18 patients, 8 had mild symptoms, 6 had moderate symptoms, and 4 had severe symptoms with age from 31 to 82 years old, 52% male and 48% female.

First, PD patients wear a shimmer sensor on the right wrist. The sensor is placed in this position to imitate the wristband that people wear in daily life (In fact, we put 5 sensors on the patient's limbs and waist as shown in Fig. 1, but in order to simulate the daily environment, we only used one sensor in this study). Then, we use the 200 Hz frequency to collect the accelerometer, gyroscope and magnetometer data of the three axes. The full scale range of the sensor is ±2.0 g, and sensitivity is 600 mV/g. A

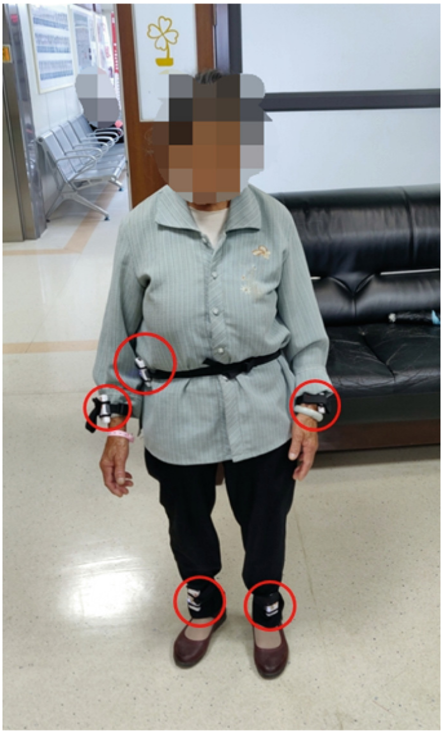


Fig. 1. Sensors wearing position of a PD patient in hospital

Lenovo ThinkPad E440 laptop connects to sensors via Bluetooth and stores data through a software called ConsensysPro.

We collected a total of 14 activities including some activities in the third part of MDS-UPDRS and some daily activities, and each action was collected from 20 s to 90 s. For daily activities, in order to reflect the differences in the habits of different patients in the real environment, we briefly tell the patients the actions that need to be performed instead of specifying the way to perform them. The activities in the third part of MDS-UPDRS can reflect the difference in the degree of patient disease, so we told the patient in detail how to perform. We described the details of all the activities in Table 1.

4 Methodology

4.1 Problem Formulation

We define activity recognition as a classification problem. Unlike the conventional classification, we will divide PD patients into two parts according to the severity of the patient. The first part simulates the previous patients, including mild, moderate and severe patients, and the other simulates the new PD patients. And then we will perform the same preprocessing and feature extraction on the data of all PD patients. Next, the classification model first learns from the previous PD patient activities data, and then

Table 1. The description of each activity

NO	Activity name	Activity description
1	Fingers tapping	Quick pinch with thumb and index finger
2	Hands closing–opening	Clench your hands and open them quickly
3	Pronation–supination	Both wrists rotate quickly to the left and right
4	Leg agility	Sitting in a chair and raising your heels repeatedly
5	Right hand flip	Continuously pat the left hand with the palm and back of the right hand
6	Left hand flip	Continuously pat the right hand with the palm and back of the left hand
7	Finger_to_nose(l)	First touch the tip of your nose with your left index finger, then touch the doctor's index finger, and repeat
8	Finger_to_nose(r)	First touch the tip of your nose with your right index finger, then touch the doctor's index finger, and repeat
9	Stand and hand raised flat	Stand and hand raised flat for 30s
10	Walking back and forth	Walk back and forth in a room
11	Free walking	Walk freely in the crowd
12	Sit-to-stand	Cross your hands in front of your chest and stand up from your seat
13	Drink water	Drink water from a cup
14	Pick up object	Pick up things from the ground

apply the learned model to identify the activities of the new PD patient, which is in line with the actual life situation. We aim to evaluate the performance of the activity recognition model when it faces PD patients in real setting and analyze the reasons behind them. Finally, we propose how to solve this problem through active learning methods.

4.2 Data Preprocessing and Feature Extraction

In order to remove the noise in the raw data, for all signal data, we use the band-pass filtering and standardization by zero-mean normalization (z-score), as Eq. (1) where μ and σ represent the mean and standard deviation of data, respectively.

Table 2. Features extracted from time domain and frequency domain

Category	Feature sets
Time domain	Mean, Standard deviation, Variance, Skewness, Kurtosis, Root mean square, Energy, Median, Range, Correlation
Frequency domain	FFT energy, Mean amplitude, Max amplitude, Spectral entropy

$$x^* = \frac{x - \mu}{\sigma} \tag{1}$$

$$SMV = \sqrt{x_i^2 + y_i^2 + z_i^2} \tag{2}$$

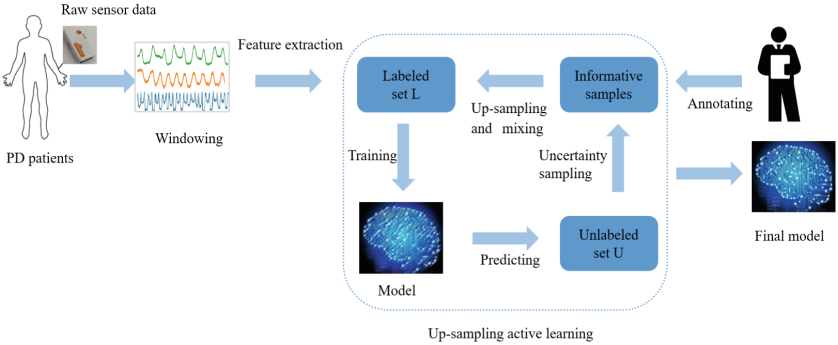


Fig. 2. Overall framework for up-sampling active learning activity recognition mode

There are 9 dimensions of raw data (a Shimmer sensor consists of three sensors and each of which has three axes of X, Y and Z). Then we got 3 extra axes after calculating signal magnitude vector (SMV) by Eq. (2), which helps to measure the intensity of activities. On each axe, as shown in Table 2, we extract independently the time domain and frequency domain features based on the sliding windows. Based on our experience, we chose a sliding window size of 3 s and the data was segmented according to the window overlap ratios of 50%. Of course, we have tried other sliding window sizes and overlap rates, but they have no significant impact on the results, and they are not the focus of our research.

4.3 Up-Sampling Active Learning

In the real setting, there are great diversities in performing a same activity between different PD patients. In this work, one of our aims is to train a robust and adaptive activities recognition model for each PD patient, which usually requires abundant labeled data with possible variations. Because annotating the activities of PD patients is time-consuming and laborious, we introduce active leaning to select the most informative samples, so as to ease the burden of activities data annotations. Besides, the proportion of activity data of a new patient is very small, which leads to model to ignore them, so up-sampling is also integrated into active learning to improve the model convergence speed.

Based on the above considerations, we propose an UAL pipeline (as illustrated in Fig. 2). We first extract features by sliding window technique from preprocessed data. Starting from an initial labeled PD patients' activity data, we iteratively update the

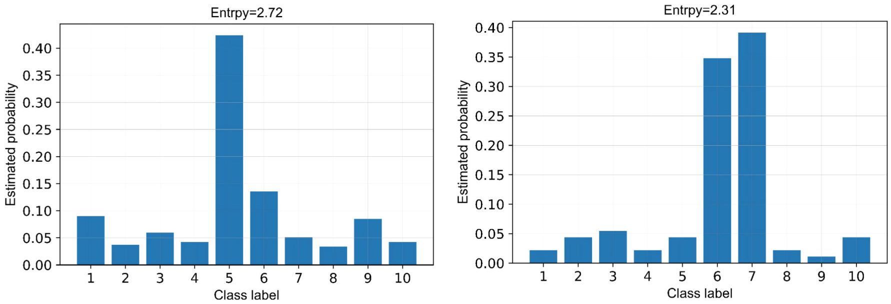


Fig. 3. The estimated probability of prediction results to two unlabeled sample and its entropy

training set by adding new patient activity samples. The steps are as follows: firstly, on previous patients dataset, we train an initial classification model using LightGBM [51]. Then the model will predict the new PD patients' activities and several informative sample will be chosen according to uncertainty sampling strategy BvSB. Next, selected samples will be passed to an oracle to be annotated. Finally, we add the newly annotated sample to previous labeled dataset after up-sampling. This iteration (the part in the dotted box in Fig. 2) will stop when a certain condition is satisfied to get the final model. The classifier LightGBM and the uncertainty sampling strategy BvSB are demonstrated as below:

LightGBM is an efficient gradient boosting decision tree algorithm (GBDT) proposed by the Microsoft team. It is an improved algorithm for GBDT and an integrated learning algorithm based on Boosting. The traditional Boosting algorithm has some drawbacks, especially in scalability and operating speed, and the emergence of LightGBM has solved these problems. Compared with other Boosting algorithms (such as XG Boost [52]), it can shorten the training time by more than 20 times without reducing the prediction accuracy, and the memory overhead is also greatly reduced. After considering the training time, memory usage and model accuracy, we chose LightGBM as the classification algorithm instead of using SVM, XGBoost, CNN, LSTM, etc., which can make our method more adaptable to the real environment.

Uncertainty sampling strategy: If we predict new patient activities using trained model by other patient activities data, we will able to select the most informative sample to annotate according the uncertainty sampling strategy. By mixing these samples with previous training set and adding them to the model to train, the decision boundary of the model will be changed, which allows the model to have better prediction results for new patient activities. Uncertainty of a prediction result can be quantified by entropy measure, least confident, and BvSB [53]. Among them, BvSB is proved to be excellent in activity recognition based on active learning [10]. Uncertainty sampling strategy using entropy measure is subject to small probability values of unimportant classes. The histogram Fig. 3 shows the estimated probability distribution of two unlabeled samples in a 10-calss classification problem, which explains why entropy cannot calculate the uncertainty well. In Fig. 3 (b), the classifier hesitates between class 6 and 7, but in Fig. 3 (a), the classifier is relatively confident in its prediction results. A prediction result like Fig. 3(a), will be considered to have higher entropy, even if the model is much more confident about the prediction of the example. In the case of multiple classifications, this situation will be even worse. To avoid above problems, we adopted the BvSB as a metrics. BvSB uses the difference between the probability values of the two classes that have the highest estimated probability value as an indicator to measure uncertainty, the sampling criterion can be described as Eq. (3), where $P(y_B|x_i, F_\theta)$ and $P(y_{SB}|x_i, F_\theta)$ denote the two highest estimated class probabilities output from classifier and u denote all unlabeled data sets. So, BvSB sampling strategy tends to select this prediction situation in Fig. 3(b).

$$x_i^{BvSB} = \underset{x_i, i \in u}{\arg\min}(P(y_B|x_i, F_\theta) - P(y_{SB}|x_i, F_\theta)) \tag{3}$$

5 Experiments Results

In this section, we evaluate and analyze the unreliability and instability of directly predicting the activity of a new patient using classifier trained by other subjects' activities dataset. Then, also on the PD patient activity data set collected in hospital, we demonstrated and analyzed the performance of active learning with up-sampling.

5.1 General Experimental Settings

All our experimental data comes from the data set mentioned in Sect. 3. After performing data preprocessing and feature extraction described in Sect. 4, we get 11115 samples from 18 PD patients. We selected 12 patients as the previous patients, and the remaining 6 patients were simulated as new patients in turn. For convenience, the classification accuracy (CA), the proportion of correctly classified samples to the number of test samples, is used as an indicator to evaluate model performance.

5.2 The Unreliability and Instability of Directly Predicting

In order to evaluate the performance of the PD patient activity recognition model trained by other subjects' activities dataset, we conducted the following experiments.

(1) As differences between patients with different symptoms, different ages etc. will cause unstable results due to the different division of the training set and test set. At the same time, in actual situations, patients are usually tested by wearing the device alone. The Leave-One-Out (LOO) validation method is explored in our experiment, which may derive more stable results and avoid over-fitting in real environment. The LOO method is a special case of cross-validation. Obviously, since there is only one way to divide m samples into m subsets (each subset contains one sample), the method is not affected by the way that how random samples are divided.

(2) Then we use Lightgbm package in the Scikit-learn library to train a PAR model eliminate randomness and select the best hyperparameters of the classification model. The learning rate is set to 0.05, the maximum number of decision trees is 300, and the model converges at 100 training epochs.

(3) We subjectively believe that different patients will have different ways of performing activities, which will make it difficult to apply an activity recognition model to all patients. But this view is ambiguous. In order to intuitively see the difference between different patients, we designed the following scheme: Firstly, we reduce the dimensions of the extracted 161-dimensional features to two dimensions through the t-SNE algorithm. Then we draw the same activity of different patients into a scatter plot based on the reduced dimensionality data. Here, for convenience of display, we only randomly selected 6 patients and 4 activities.

5.3 Results and Discussions on Cross-Subject Prediction

Table 3 exhibits the classification accuracy of the model trained by previous dataset for each new patient's activity. The table suggests: 1) If the model is trained by previous data but apply to never seen new patient data, the prediction results will be relatively poor and unstable. It is worth mentioning that the model has a classification accuracy of 99% on the test set of the previous data. 2) The model has different classification accuracy on different patients, the highest is 90.1%, but the lowest is only 44.3% recognize the PAs of patients with severe PD.

The main reason for the large fluctuations in the performance of the model among different people is that different patients have different symptoms, severity, and activity habits. These factors firstly lead to differences in the data collected by the sensors, leading to different feature spaces of the data. If a new sample is outside the decision boundary learned by the classifier, it will be difficult for the classifier to make a correct judgment. The classifier has a 90.1% accuracy for patient 4, which is probably because the patient's activity execution method is very similar to some patients in the training set. However, patient 6 was just the opposite of the above, so the classification accuracy is only 44.3%.

Through the visualization results shown in Fig. 4, we can better illustrate this problem, where the horizontal and vertical coordinates are the data obtained after dimensionality reduction of 161 features and the different colors indicate the data of different patients. It can be seen from Fig. 4 that different patients gather in different areas on a two-dimensional plane, especially the three activities of "finger tapping", "sit-to-stand", and "drink water". Take activity "free walking" as an example. If the model training set only includes the data of patients p1 to p5 (The part in the dotted box in the Fig. 4, then the data of patient p6 will be outside the decision boundary, and the classifier will have difficulty making accurate predictions.

Carefully observe the data distribution of "free walking", we can find that its spatial distribution of activity features is relatively chaotic. The data of patients p1, p3, p4, and p5 are clustered together, which shows that they perform this activity very similarly. However, patients p2 and p6 were gathered in other areas, which shows that they are very different from others in performing this activity. This is mainly because patients p1, p3, p4, and p5 are patients with mild or moderate PD, but patients p2 and p6 are patients with severe PD. And the mobility of severe patients is severely restricted, so they cannot complete this activity well. It is worth mentioning that even if patients 2 and 6 are both severe PD patients, their data distribution is still quite different, which

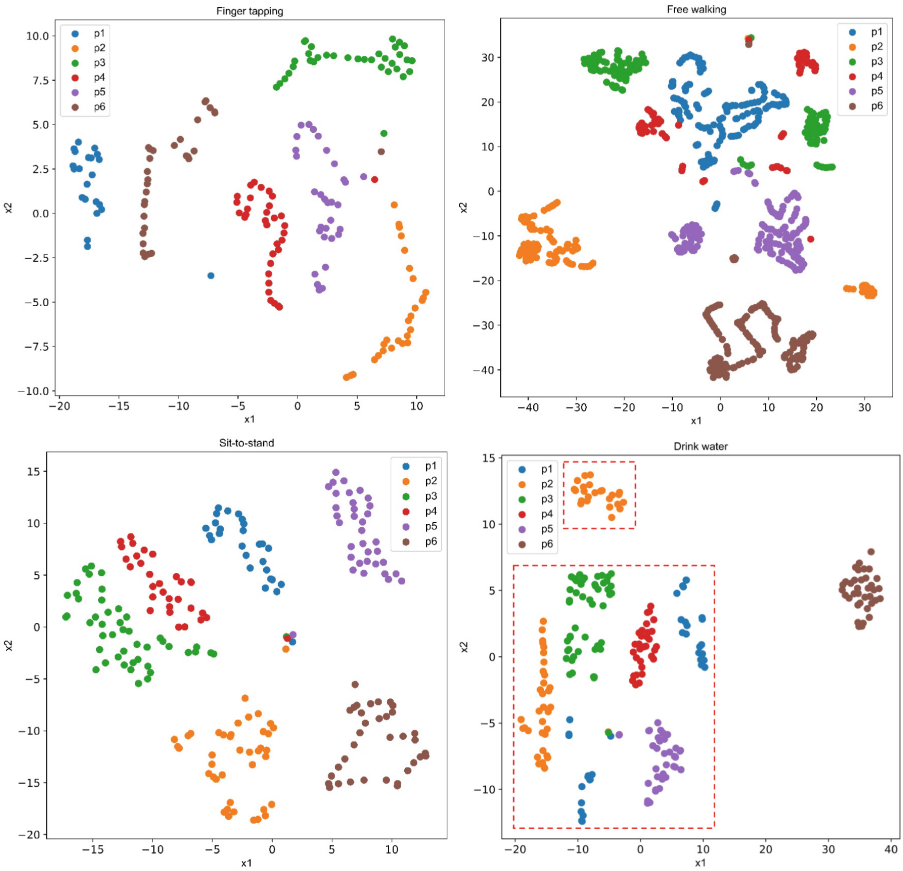


Fig. 4. Feature distribution of 4 kinds of activities in two-dimensional space

explains why the training set contains data from severe PD patients, but the classifier still unable to accurately.

Table 3. The classification accuracy to 6 new patients with the model trained by the previous 12 patients

Patient ID	1	2	3	4	5	6
Severity of illness	moderate	severe	mild	mild	mild	severe
Accuracy (%)	66.6 ± 1.6	59.1 ± 0.7	74.5 ± 1.0	90.1 ± 1.9	64.3 ± 1.1	44.3 ± 0.5

Table 4. The number of labeled samples and the proportion of the total samples required to achieve 99% accuracy with different methods

Patient	Method			
	UAL(Ours)	BvSB	entropy	random
Patient 1	**70 (19.18%)**	105 (28.77%)	90 (24.66%)	340 (93.15%)
Patient 2	**75 (15.96%)**	100 (21.28%)	105 (22.34%)	365 (77.66%)
Patient 3	**70 (15.91%)**	80 (18.18%)	105 (23.86%)	195 (44.32%)
Patient 4	**20 (5.63%)**	25 (7.04%)	35 (9.86%)	165 (47.14%)
Patient 5	**85 (20.48%)**	125 (30.12%)	175 (42.17%)	260 (63.41%)
Patient 6	**105 (25.0%)**	135 (32.14%)	170 (40.48%)	305 (72.62%)

Table 5. The highest classification accuracy value (the first number) that each method can achieve and the proportion of the total samples required to achieve corresponding highest accuracy (the second number)

Patient	Method			
	UAL(Ours)	BvSB	entropy	random
Patient 1	**100% (20.5%)**	100% (41.1%)	100% (34.2%)	100% (100.0%)
Patient 2	**99.51% (22.3%)**	**99.61% (23.4%)**	99.51% (26.6%)	99.51% (78.7%)
Patient 3	**100% (40.9%)**	100% (55.7%)	100% (68.2%)	100% (62.5%)
Patient 4	**100% (8.5%)**	**100% (8.5%)**	100% (25.4%)	100% (85.7%)
Patient 5	**100% (28.9%)**	100% (69.9%)	100% (75.9%)	99.78% (85.4%)
Patient 6	**99.89% (51.2%)**	99.78% (66.7%)	99.78% (57.1%)	99.57% (95.2%)

5.4 Active Learning with Up-Sampling

In order to verify and evaluate the efficiency of our proposed method and the effectiveness of the up-sampling, we conducted the following experiments. Firstly, 70% of the data of each new patient is used as a candidate set, and the remaining 30% is used to test the accuracy on the new data set. Then, we train a classifier by previous patients training set. When the model converges, we make predictions on the test set and candidate set at the same time, and record the accuracy of the model on the test set. By calculating class probabilities of samples in the candidate set, we select samples for annotation according to the sampling strategy. Next, we copy the newly annotated data into the same five copies and mix them all into the previous training set for the classifier retraining in next iteration. It is worth noting that, after comprehensively considering the calculation amount and performance of the model, we select 5 samples for annotation in each iteration. We compare the performance of three different sampling strategies, BvSB, entropy and random, and compare the effect of up-sampling on the results.

5.5 Results and Discussions on Active Learning with Up-Sampling

Figure 5 presents the curves of the classification accuracy values of 6 new patients as a function of the number of labelled samples using three different sampling policies, where the x-axis is the number of annotated samples and the y-axis is the classification accuracy of the classifier on the test set.

The number of labeled samples and the proportion of the total samples required to achieve corresponding highest accuracy are listed in Table 5. The best results in Table 4 and 5 are shown in bold. From Fig. 5, Table 4 and 5, we can see that:

1) For all patients, our proposed UAL method can achieve 99% accuracy with the smallest number of labeled samples, followed by active learning methods based on BvSB and entropy, and the worst random sampling. We tend to randomly select some samples for labeling when active learning is not applied, which is often very inefficient. We attribute this result as follows: Firstly, the BvSB sampling strategy allows the model to quickly and accurately refine the decision boundary by selecting the most informative samples, which allows the model to quickly adapt to samples of new patients and make high-precision predictions. Secondly, through the up-sampling, the weight of the newly labeled data in the model can be increased, which allows the model to converge more quickly.
2) The sampling strategy based on BvSB tends to improve the accuracy faster than the sampling strategy based on entropy. It's because BvSB strategy is inclined to choose samples meeting Eq. (3) in Sect. 4, such samples are just on the decision boundary of the two classes. But the entropy sampling scheme tends to select samples whose estimated probabilities are scattered over all classes with similar probability values, such samples may not have high uncertainty as described in Sect. 4.
3) For patients 2 and 6, when we didn't annotate their activities data, the classification accuracy was only 59.1% and 44.3% respectively. Because they are patients with severe PD and exhibits symptoms rigidity and bradykinesia, they are very slow to perform an action, which is very different from patients with mild PD. For example, patient 4, he is a mild PD patient, he performs these activities smoothly and standardly, which is the same as most people in the training set, so the classifier's classification accuracy for him is 90% at the beginning. But with our method UAL, the classifier only needs 20 additional labeled samples to achieve a classification accuracy of 99%, which fully demonstrates the efficiency of UAL.
4) It can be seen from Table 5 that for patients 1, 3, 4, and 5, all methods except random sampling can achieve 100% classification accuracy, but our proposed method requires the least number of labeled samples. For patients 2 and 6, our method UAL and BvSB method are almost the same in accuracy, but our method requires fewer labeled samples. The difference between UAL and BvSB is that we have added up-sampling to increase the weight of the data in the training process, which allows the model to improve accuracy faster.

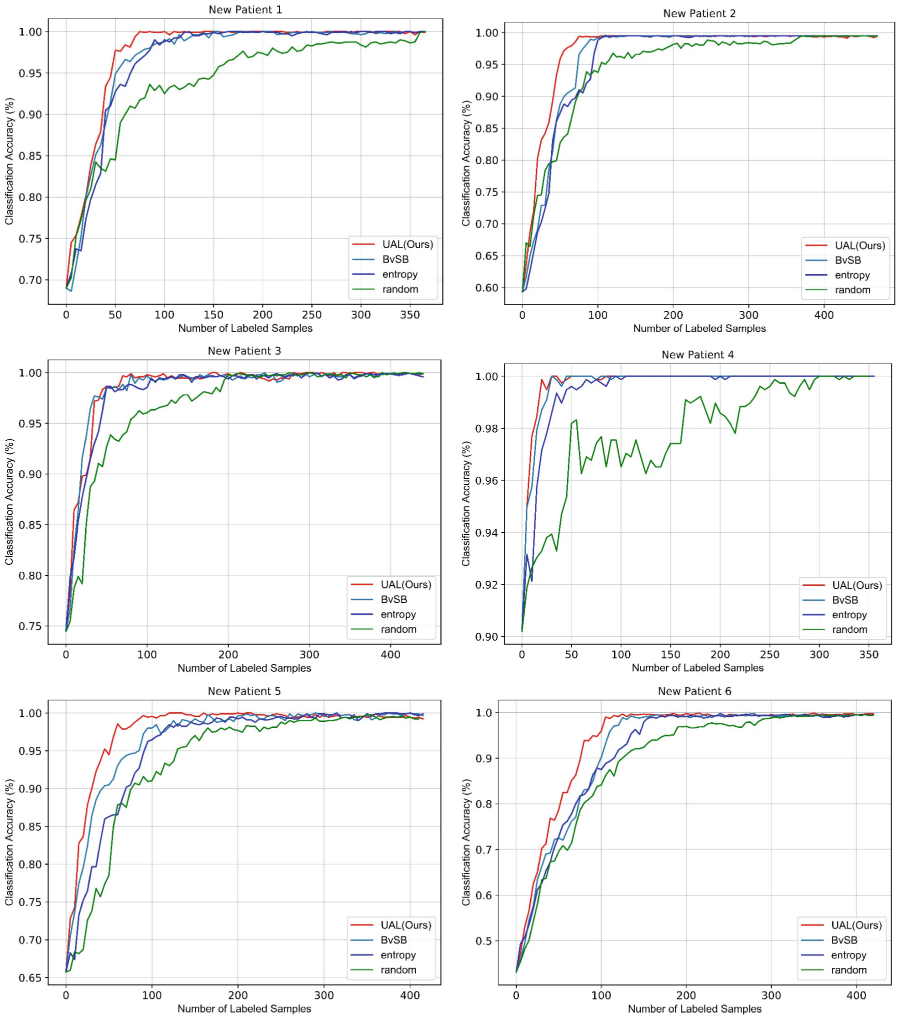


Fig. 5. The curves of the classification accuracy values of 6 new patients as a function of the number of labelled samples

6 Conclusions

In this paper, we analyzed the difficulties in identifying the activities of PD patients. An up-sampling active learning method was proposed for effective long-term monitoring of PD patients. This method iteratively selects the most informative samples for labeling, and then adds them to the training set after up-sampling. We conducted experiments on the data set collected from the local hospital to evaluate the performance of UAL. The experimental results show that the effect of UAL is different for patients with different symptoms. For the 6 new PD patients after annotating 5.63% to 25.00% of the data, the model accuracy can reach up to 99%. Among them, for severe PD patients, this method has a more obvious performance improvement and need more labeled data. Compared with other active learning sampling strategies, the UAL method that combines BvSB sampling strategy and up-sampling trick can converge to the better accuracy, which represents a lower annotation cost. In future work, we plan to combine active learning and transfer learning to further reduce annotation costs and improve the robustness of the model, and analyze the progress of the patient's condition through the patient's activity data.

References

1. Jankovic, J.: Parkinson's disease: clinical features and diagnosis. **79**(4), 368–376 (2008)
2. Dorsey, E.R., et al.: Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. **68**(5), 384–386 (2007)
3. Nutt, J.G., Gancher, S.T., Woodward, W.R.: Does an inhibitory action of levodopa contribute to motor fluctuations? **38**(10), 1553–1553 (1988). JN
4. Merello, M., Lees, A.J.: Beginning-of-dose motor deterioration following the acute administration of levodopa and apomorphine in Parkinson's disease. **55**(11), 1024–1026 (1992). JoN, Neurosurgery, Psychiatry
5. Maetzler, W., Klucken, J., Horne, M.: A clinical view on the development of technology-based tools in managing Parkinson's disease. JMD **31**(9), 1263–1271 (2016)
6. Albani, G., et al.: An integrated multi-sensor approach for the remote monitoring of Parkinson's disease. **19**(21), 4764 (2019)
7. De Pessemier, T., Martens, L.: Heart rate monitoring, activity recognition, and recommendation for e-coaching. Multimed. Tools Appl. **77**(18), 23317–23334 (2018). https://doi.org/10.1007/s11042-018-5640-2
8. Ryder, J., Longstaff, B., Reddy, S., Estrin, D.: Ambulation: a tool for monitoring mobility patterns over time using mobile phones. In: 2009 International Conference on Computational Science and Engineering, 2009, pp. 927–931. IEEE (2009)
9. Emmanouil, G., et al.: MyHealthAvatar: personalised and empovermnet health services through internet of things technologies. In: 2014 4th International Conference on Wireless Mobile Communication and Healthcare Transforming Healthcare through Innovatins in Mobile and Wireless Technologies (MOBIHEALTH), 2014, pp. 331–334. IEEE (2014)
10. Bi, H., Perello-Nieto, M., Santos-Rodriguez, R., Flach, P.: Human activity recognition based on dynamic active learning. IEEE J. Biomed. Health Inform. **25**(4), 922–934 (2020). JIJoB, Informatics H
11. Qi, J., Yang, P., Waraich, A., Deng, Z., Zhao, Y., Yang, Y.: Examining sensor-based physical activity recognition and monitoring for healthcare using Internet of Things: a systematic review. **87**, 138–153 (2018)
12. Zhang, M., Sawchuk, A.A.: USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 1036–1043 (2012)
13. Twomey, N., et al.: The SPHERE challenge: activity recognition with multimodal sensor data (2016). arXiv preprint arXiv:1603.00797
14. Martín, H., Bernardos, A.M., Iglesias, J., Casar, J.: Activity logging using lightweight classification techniques in mobile devices. **17**(4), 675–695 (2013)
15. Kwapisz, J.R., Weiss, G.M., Moore, S.: Activity recognition using cell phone accelerometers. **12**(2), 74–82 (2011)

16. Xu, H., Pan, Y., Li, J., Nie, L., Xu, X.I.: Activity recognition method for home-based elderly care service based on random forest and activity similarity. IEEE Access **7**, 16217–16225 (2019)
17. Cook, D.J., Krishnan, N.C., Rashidi, P.J.: Activity discovery and activity recognition: a new partnership. 43 (3), 820–828 (2013)
18. Khan, A.M., Tufail, A., Khattak, A.M., Laine, T.: Activity recognition on smartphones via sensor-fusion and kda-based svms. **10**(5), 503291 (2014)
19. Ouchi, K., Doi, M.: Indoor-outdoor activity recognition by a smartphone. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 600–601 (2012)
20. Ha, S., Choi, S.: Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 381–388. IEEE (2016)
21. Bianchi, V., Bassoli, M., Lombardo, G., Fornacciari, P., Mordonini, M., De Munari, I.: IoT wearable sensor and deep learning: an integrated approach for personalized human activity recognition in a smart home environment. **6**(5), 8553–8562 (2019)
22. Mutegeki, R., Han, D.S.: A CNN-LSTM approach to human activity recognition. In: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 362–366. IEEE (2020)
23. Alawneh, L., Mohsen, B., Al-Zinati, M., Shatnawi, A., Al-Ayyoub, M.A.: Comparison of unidirectional and bidirectional LSTM networks for human activity recognition. In: 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 1–6. IEEE (2020)
24. Chen, L., Hoey, J., Nugent, C.D., Cook, D.J., Yu, Z.J.: Sensor-based activity recognition. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **42**(6), 790–808 (2012)
25. Thomaz, E., Essa, I., Abowd, G.D.: A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1029–1040 (2015)
26. Merck, C.A., Maher, C., Mirtchouk, M., Zheng, M., Huang, Y., Kleinberg, S.: Multimodality sensing for eating recognition. In: PervasiveHealth, pp. 130–137 (2016)
27. Qi, J., Yang, P., Min, G., Amft, O., Dong, F., Xu, L.: Advanced internet of things for personalised healthcare systems: a survey. Pervasive Mob. Comput. **41**, 132–149 (2017)
28. Peng, X., Yang, Y., Wang, X., Li, J, Qi, J., Yang, P.: Experimental analysis of artificial neural networks performance for accessing physical activity recognition in daily life. In: 2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), pp. 1348–1353. IEEE (2020)
29. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: A review of human activity recognition methods. Front. Robot. AI **2**, 28 (2015). JFiR, AI
30. Krishnan, N.C., Cook, D.J.: Activity recognition on streaming sensor data. Pervasive Mob. Comput. **10**, 138–154 (2014)
31. Zhang, M., Sawchuk, A.A.: Human daily activity recognition with sparse representation using wearable sensors. **17**(3), 553–560 (2013). JIJOBInformatics H
32. Basilakis, J., Lovell, N.H., Redmond, S.J., Celler, B.G.: Design of a decision-support architecture for management of remotely monitored patients. IEEE Trans. Inf. Technol. Biomed. **14**(5), 1216–1226 (2010)
33. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (eds.) Pervasive Computing. Pervasive 2004. LNCS, vol. 3001, pp. 1–17. Springer, Berlin, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24646-6_1

34. Cheng, W.Y., Scotland, A., Lipsmeier, F., Kilchenmann, T., Jin, L., Schjodt-Eriksen, J., et al: Human activity recognition from sensor-based large-scale continuous monitoring of Parkinson's disease patients. In: 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), pp. 249–250 (2017)
35. Albert, M.V., Toledo, S., Shapiro, M., Kording, K.: Using mobile phones for activity recognition in Parkinson's patients. Front. Neurol. **3**, 158 (2012)
36. Bot, B.M., et al.: The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci. Data **3**(1), 1–9 (2016)
37. Kazemimoghadam, M.: Fey NP an activity recognition framework for continuous monitoring of non-steady-state locomotion of individuals with Parkinson's disease. Appl. Sci. **12**(9), 4682 (2022)
38. Kaur, S., Aggarwal, H., Rani, R.: Diagnosis of Parkinson's disease using deep CNN with transfer learning and data augmentation. Multimed. Tools Appl. **80**(7), 10113–10139 (2020). https://doi.org/10.1007/s11042-020-10114-1
39. Balabka, D.: Semi-supervised learning for human activity recognition using adversarial autoencoders. In: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, pp. 685–688 (2019)
40. Ma, Y., Ghasemzadeh, H.: Labelforest: non-parametric semi-supervised learning for activity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 01. Pp. 4520–4527 (2019)
41. Qin, X., Chen, Y., Wang, J., Yu, C.: Cross-dataset activity recognition via adaptive spatial-temporal transfer learning. **3**(4), 1–25 (2019)
42. Wang, J., Zheng, V.W., Chen, Y., Huang, M.: Deep transfer learning for cross-domain activity recognition. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **3**(4), 1–25 (2019)
43. Khan MAAH, Roy N, Misra A: Scaling human activity recognition via deep learning-based domain adaptation. In: 2018 IEEE international conference on pervasive computing and communications (PerCom),. IEEE, pp 1–9 (2018)
44. Stikic, M., Van Laerhoven, K., Schiele, B.: Exploring semi-supervised and active learning for activity recognition. In: 2008 12th IEEE International Symposium on Wearable Computers, pp 81–88. IEEE (2008)
45. Liu, R., Chen, T., Huang, L.: Research on human activity recognition based on active learning. In: 2010 International Conference on Machine Learning and Cybernetics, pp. 285–290. IEEE (2010)
46. Diethe, T., Twomey, N., Flach, P.A.: Active transfer learning for activity recognition. In: ESANN (2016)
47. Hoque, E., Stankovic, J.: AALO: activity recognition in smart homes using active learning in the presence of overlapped activities. In: 2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops, 2012, pp. 139–146. IEEE (2012)
48. Hossain, H.S., Al Haiz Khan, M.A., Roy, N.: DeActive: scaling activity recognition with active deep learning. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **2**(2), 1–23 (2018)
49. Wang, D., Shang, Y.: A new active labeling method for deep learning. In: 2014 International Joint Conference on Neural Networks (IJCNN), pp. 112–119. IEEE (2014)
50. Zhou, S., Chen, Q., Wang, X.: Active deep networks for semi-supervised sentiment classification. In: Coling 2010: Posters, pp. 1515–1523 (2010)
51. Ke, G., et al.: Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. **30**, 3146–3154 (2017)

52. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
53. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2372–2379. IEEE (2009)

Your Day in Your Pocket: Complex Activity Recognition from Smartphone Accelerometers

Emma Bouton-Bessac[1,2](✉), Lakmal Meegahapola[1,2],
and Daniel Gatica-Perez[1,2]

[1] Idiap Reseach Institute, Martigny, Switzerland
emma.bouton@idiap.ch
[2] EPFL, Martigny, Switzerland

Abstract. Human Activity Recognition (HAR) enables context-aware user experiences where mobile apps can alter content and interactions depending on user activities. Hence, smartphones have become valuable for HAR as they allow large, and diversified data collection. Although previous work in HAR managed to detect simple activities (i.e., sitting, walking, running) with good accuracy using inertial sensors (i.e., accelerometer), the recognition of complex daily activities remains an open problem, specially in remote work/study settings when people are more sedentary. Moreover, understanding the everyday activities of a person can support the creation of applications that aim to support their well-being. This paper investigates the recognition of complex activities exclusively using smartphone accelerometer data. We used a large smartphone sensing dataset collected from over 600 users in five countries during the pandemic and showed that deep learning-based, binary classification of eight complex activities (sleeping, eating, watching videos, online communication, attending a lecture, sports, shopping, studying) can be achieved with AUROC scores up to 0.76 with partially personalized models. This shows encouraging signs toward assessing complex activities only using phone accelerometer data in the post-pandemic world.

Keywords: smartphone sensing · human activity recognition · accelerometer · deep learning

1 Introduction

In Human Activity Recognition (HAR), various human activities such as walking, running, sitting, [...], cooking, driving are recognized. The data can be collected from wearable sensors or accelerometer or through video frames or images [9]. HAR is possible thanks to sensor data from modalities such as accelerometer, gyroscope, or location [16,20]. According to Plötz et al. [16], the main challenges of HAR are the lack of data and the poor quality and labeling of the data. Recent devices like smartwatches allow for good-quality data for HAR. For

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 247–258, 2023.
https://doi.org/10.1007/978-3-031-34586-9_17

example, using a smartwatch, Laput et al. [10] obtained high accuracies for classifying 25 complex hand activities. However, smartwatch adoption is much lower compared to smartphones, and according to Coorevits et al. [5], most people tend to stop using smartwatches and wearables after six months of use.

Using smartphones for HAR seems promising given their ubiquity: more than 80% of people own a smartphone, which could simplify data collection and increase the amount of data. Data collection can be performed on diverse populations, and continuous collection is possible. The data collected are diverse because there are numerous sensors in a smartphone, such as an accelerometer, gyroscope, light sensor, magnetic field, app usage, typing and touch events, etc. [14]. Multiple sensing modalities also allow recognizing complex activities such as eating [13] and drinking [4], and even complex psychological states such as mood [18]. Using smartphones for young adults' well-being is also increasingly popular [14] because of the high smartphone ownership in this population. Understanding one's everyday activities can help create applications to improve mental health. Also, using data from different countries involves taking into account different cultures, people, sensor qualities, and ways to carry a smartphone (pocket, backpack, purse, etc.). Therefore, data from multiple countries should generalize better, although bringing additional challenges. Using multiple sensing modalities, while informative, could be costly in terms of battery life. Hence, there is a push towards only using low-cost inertial sensors for HAR [2].

Previous work on HAR that use inertial sensors focuses on inferring relatively simple activities such as walking, sitting, climbing stairs, and sleeping [3,8]. However, recognizing complex activities can be helpful in various situations, such as elderly care and patient tracking [10,17] and for habit tracking (e.g., to help people quit smoking [10]). Moreover, due to the pandemic, most people's everyday life has changed to a more sedentary lifestyle, making the HAR tasks even more challenging because the informativeness of smartphone accelerometers could be less.

In this work, we attempt to address the research question (RQ): Can only raw accelerometer data be used to recognize complex daily activities with data collected during the pandemic (remote study setting)? In addressing this RQ, two contributions are provided:

Contribution 1: We examine a real-life smartphone sensing dataset that contains over 216K self-reports from 637 college students in five countries. The dataset was collected for four weeks during the pandemic. We perform a descriptive data analysis to identify the most common complex activities reported by participants.

Contribution 2: We define and evaluate binary inference models for eight complex daily activities: Sleeping, Eating, Studying, Attending a lecture, Online Communication and Social Media, Watching videos or TV, Sports, and Shopping, all of which represent facets of the everyday life of young adults. Using only raw accelerometer data and deep learning, we show that AUROC scores in the range of 0.51–0.62 can be achieved with population-level models, and it could be improved to AUROC scores in the range of 0.56–0.76 with hybrid models.

To the best of our knowledge, our work contributes to understanding how the sole use of smartphone accelerometer data can be used for the inference of complex activities like the ones we study here. The pandemic context enhanced remote work and sedentary lifestyles, so it is a setting worth investigating. The paper is organized as follows. In Sect. 2, the related work is presented. Then, the methods and results are explained in Sect. 3 and Sect. 4 respectively. Finally, the main findings are discussed in Sect. 5, and the paper is concluded in Sect. 6.

2 Background and Related Work

2.1 Smartwatches and HAR

Wearables for HAR. Laput et al. [10] managed to capture fine-grained hand activities using smartwatches. There were 25 hand activities such as clapping, drinking, or door opening. Using Fast Fourier Transform and Convolutional Neural Networks, the method yielded 95.2% accuracy across the 25 hand activities. This work could be used to track habits such as smoking or for eldercare monitoring systems. One disadvantage is that the user must wear the device on their active arm, whereas smartwatches are usually worn on the passive arm. HAR with smartwatches on the passive arm would be more challenging but more adapted to real life. Another challenge with smartwatches, shown by Straczkiewicz et al. [21], is that about 15.6% people do not follow the data collection protocol regarding smartwatch placement, such as wearing the watch on the 'wrong' arm. This study shows that the data collection for HAR can be very challenging, as simply wearing a sensor on the non-ideal arm can decrease performance. These results are also valid for real-life applications: if the sensors are misplaced, the detected activity could be incorrect.

Smartphones vs. Smartwatches for HAR. Raihani et al. [15] showed that classifiers can perform as well as when the accelerometer is placed in the pocket rather than on the wrist for basic activities (sitting, walking, running). Smartphones have a practical advantage in the long run, as many users stop using their smartwatches after a few months [5]. Performing HAR with smartphones can be as efficient as with wearables, and more data can be used as the ownership of smartphones is higher than that of smartwatches. Furthermore, combining smartphone sensors and wrist-worn motion sensors is even more effective than only using smartwatches [19]. Such work evaluated basic activities along with more complex ones like smoking, biking, or drinking coffee. The results showed that combining the sensors from the phone and the watch improves the performance by 21%, for an overall F1 measure of 96%. However, the work in [19] was performed in a lab setting, and its application to real-life cases will probably yield lower performance.

2.2 Smartphones and HAR

Sensors and Features. Smartphone sensors can be used to infer a variety of human activities and states. For example, mood can be inferred from social interaction data [11]. Features like the number of SMS, emails, and apps used are fed into various machine learning models to assess user mood. The method achieved 93% accuracy after a two-months personalized training period. Guvensan et al. [7] used the smartphone's accelerometer, gyroscope, and magnetometer sensors to assess the transport mode. The method achieved 95% accuracy with supervised learning approaches. Hassan et al. [8] extracted features (mean, frequency skewness, average energy) from the smartphone's gyroscope and accelerometer and fed them into a Deep Belief Network. The method achieved 89.61% accuracy on basic activities (walking, sitting, walking upstairs/downstairs) and the transitions between two activities. Wu et al. [22] used the same activities performed at different paces and collected data from the accelerometer. Their method obtained accuracies between 52–79% for stair walking and up to 100% for sitting; adding the gyroscope data improved the performance by 3.1–13.4%. Della Mea et al. [12] used the smartphone's accelerometer to infer household activities such as working at the computer, ironing, or sweeping the floor. Their proposed method obtained an accuracy above 80%, even when the phone was in the pocket. This gives initial evidence to support the hypothesis that the phone's accelerometer alone could also be used for recognizing complex activities.

Complex Activities. Ranasinghe et al. [17] defined a complex activity as a succession of simple actions. The actions are composed of operations, which are the basic steps constituting the actions. For instance, the complex activity "Party" can be broken down into actions such as "meet with friends", "enter a bar", and "order a drink". These actions can then be broken down into operations like "push the door handle" or "grab the glass". Complex activities can include interactions with objects or individuals (such as eating, communicating online, and partying) and last longer in time. HAR can monitor the complex activities of elderly people and improve their quality of life. Healthcare monitoring applications are also an interesting field, and using only the smartphone to recognize activities is not invasive, compared to previous work that often uses body sensors [23]. Using the minimum amount of sensors allows for a spare battery and would also be more efficient memory-wise. However, it is more challenging because there will be less data, and this data can be less meaningful for some complex activities.

Our work differs from previous work regarding the inferred activities and the sensors used. We aim to infer complex activities like studying or eating, exclusively using raw accelerometer data collected in everyday life. This makes the inference challenging compared to HAR models trained with data collected in in-lab settings. Further, as mentioned in Sect. 3, the dataset being collected during the COVID-19 pandemic represents a challenge because accelerometer data will likely be similar for different activities. Hence, there is a novelty in studying how complex activity recognition models perform with data collected during the pandemic.

3 Methods

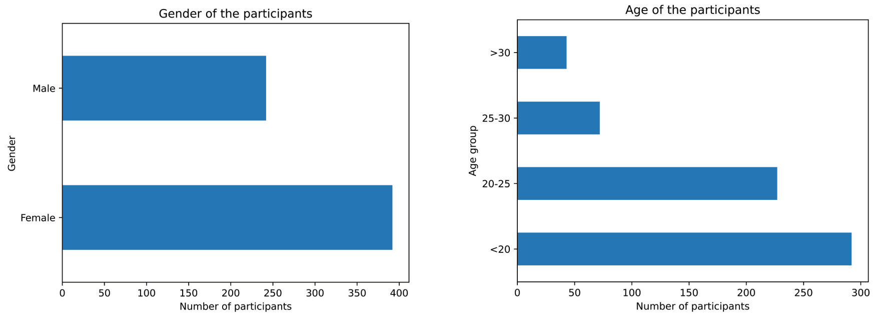
3.1 Dataset

The anonymized data used in this study was collected as part of a European Union Horizon 2020 Project called WeNet [6]. The data were collected in the fall of 2020. The original study aimed to measure aspects of the diversity of university students based on social practices and related daily behaviors, combining mobile surveys and smartphone sensor data. The study was conducted at Aalborg University (Denmark), the London School of Economics (United Kingdom), the National University of Mongolia (Mongolia), Universidad Católica "Nuestra Señora de la Asunción" (Paraguay), and the University of Trento (Italy).

A sample of volunteer students participated in a four-week data collection. The students were approached by the data collectors via an email to the entire population enrolled in the universities that took part in the survey [1]. After having consented to the processing of their personal data, agreed to participate and have consented to be contacted along with having a smartphone version of Android 6.0 or higher, participants filled out a time diary via a mobile app [1]. Participants were 61% females and average age was 22 years old (see Fig. 1). The app sent notifications every hour for the four weeks, asking the participant to complete a time diary (also referred to as self-report) to report their current activity, among other variables not used in this paper. If the participant could not answer the questionnaire, they could fill it in later (for example, when they woke up, they could indicate they have been sleeping for the past hours). The students received incentives at the end of the study. The activity list was defined according to previous survey work in sociology. In the meantime, the application collected data from 34 sensors, such as the accelerometer, gyroscope, battery level, app usage, etc. Here, only the raw accelerometer data will be used (other sensor data could have been used, but they were not considered as our focus here is specifically on the accelerometer data). After data pre-processing and filtering, approximately 40K self-reports were available for analysis.

3.2 Data Preparation

Class Selection. Figure 2a shows the dataset's number of events per activity. The large class imbalance is not surprising, but it means that there is not enough data for all activities: *Travelling* counts only 19 events across all five countries, which is not enough for a model to learn. Therefore, *Movie, theatre, concert*, *Hobbies*, *Arts*, *Happy hour/drinking*, *Other entertainment*, *Entertainment Exhibit, Culture*, and *Travelling* have not been taken into account in this work, because of the lack of data. Other activities were not considered because they are too broad: *Personal care*, *Games*, *Social life*, or *Voluntary work*, involve many possible complex activities. Activities like *Nothing special*, *Break*, and *Other* are too general and thus not interesting to infer. In addition, we decided to merge some classes: *Shopping* and *Other shopping* were merged into *Shopping*; while *Calling*,

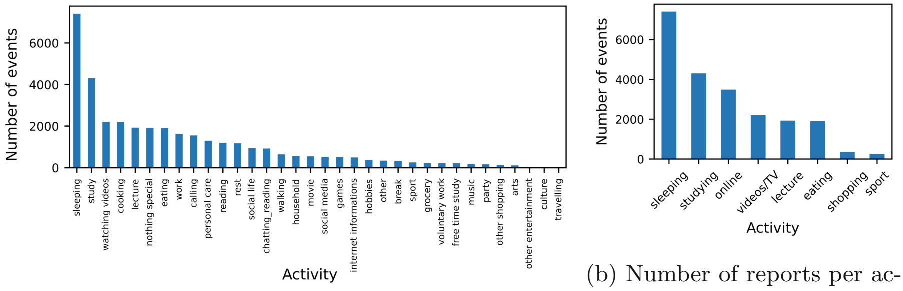


(a) Gender distribution of the dataset. (b) Age distribution of the dataset.

Fig. 1. Gender and age distribution of the dataset.

Chatting/reading, *Reading internet information*, and *Social media* were merged into *Online communication and social media.*

After this process, eight classes were finally kept: *Sleeping, Eating, Studying, Attending a lecture, Online communication and social media, Watching videos/TV, Sports, Shopping*. Their distribution is shown in Fig. 2b. These eight classes are still diverse and specific enough to allow training. In addition, they are interesting cases of the complex everyday activities of university students. Many of them directly impact health (sleeping, eating, doing sports) and some indirectly (Online communication and social media, studying), so these activities are worth inferring regarding young adults' well-being.



(a) Number of reports per activity for the whole dataset. (b) Number of reports per activity for the final list of activities.

Fig. 2. Number of reports per activity.

Pre-Processing. Previous HAR work has used time windows of 2–30 s to infer basic activities [3, 19]. However, the activities of interest in this work are complex, which means that they last longer [17]; thus, it might be harder to recognize complex activities in only a couple of seconds. Another aspect of this task is that it is unclear exactly when the participant is performing the reported activity. As they fill out the form every 30 min, it is unclear whether they are doing the activity during the entire 30-min period, only before the report or only after. Therefore, one assumption was that the user performs the activity sometime during the 3 min before the self-report time (i.e., the timestamp when the participant completed the self-report). It was further assumed that the person did not perform the reported activity during the completion of the report itself (i.e., running while completing a report about running), so the data corresponding to this specific time has been removed. Hence, the following process was applied to the data:

1. For each user's self-report, select the accelerometer data for the last 3 min before the report time.
2. Remove the time during which the user fills the self-report.
3. Re-sample the accelerometer data to the average sampling frequency of the dataset (3.33 Hz)
4. If a report is shorter than 600 samples (3 min * 60 s * 3.33), the data point is discarded.

Upon inspection, we noticed significant missing data: even though the data pre-processing was the same for all five countries, many events were discarded because there needed to be accelerometer data within a 10-min time window around the self-report time. For Mongolia and Italy, most reports were discarded because there was no data. Paraguay, Denmark, and the UK have a small number of empty reports. Mongolia and Italy are where most data was gathered, so the data loss is significant: they represent 160K reports, whereas Paraguay, Denmark, and the UK gather 31K reports. This represents a challenge because it reduces the amount of data available for deep learning. The remaining self-reports were resampled to the average sampling frequency of the dataset (see Fig. 4). The average sampling time was computed for seven users of each country and rounded to 300 ms.

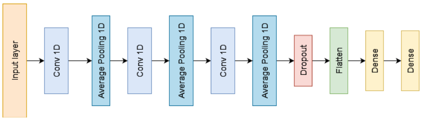


Fig. 3. Model Architecture.

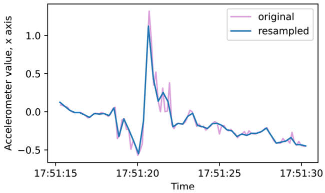


Fig. 4. "Attending lecture" accelerometer value for a Denmark participant, x axis.

3.3 Deep Learning

The previous section discussed the last three minutes of accelerometer data before each self-report was used. The input data has a shape of 3 × 600: three accelerometer axes, x, y, and z, and a data length of 600. Each report is labeled with the corresponding activity. The data is then fed to a Deep Learning model. Several architectures were implemented, using 1D Convolutional layers or LSTM layers, and the best performing model was used, as shown in Fig. 3. 1D Convolutional layers were used because the input data is a sequence. LSTM models were explored because it is usually used to process sequences of data. It did not give better results than 1D Convolutions in this case. In addition, the Adam optimizer and binary cross-entropy were used. The performance measures used were accuracy, the area under the receiver operating characteristic curve (AUROC), and the F1 score. We reported all three metrics because accuracy makes sense in a balanced class setting. However, AUROC and F1 scores with macro averaging make sense in the imbalanced class setting because they give equal emphasis to both majority and minority classes. The evaluation was done with 10-fold cross-validation.

The binary classification was performed for each class, such as ‘sleeping’ and ‘not sleeping’. For each training split, 60% of samples from the positive class were randomly selected for training, 20% for validation, and 20% for testing. Two settings were tested, with balanced and imbalanced data. The same amount of samples were selected from the seven other classes for the balanced case. All data from the seven other classes were used for the imbalanced case. Training on an imbalanced dataset was done to determine whether niche events (events that do not occur often) can be inferred. For instance, shopping is a fraction of the week of a student, and it is interesting to see whether it could be detected.

The training was run on a population level, meaning that data is split users-wise: users can only be in one set (training users are not in validation or test). The results of this experiment will show whether complex activity recognition can perform well on new users. Another experiment was to split the data samples-wise: reports were randomly split into train, validation, and test.

4 Results

4.1 Population Level Model

The results for the population-level model can be seen on Table 1.

Balanced Dataset. This could be considered as the base-case accuracy without any personalization. Hence, results indicate that it is likely every user’s smartphone usage is very different, and the model does not perform well on a new user and needs personalizing. Sleeping has the highest AUROC score of 0.62. The reason could be that it is the most significant class, so there is more data to train the model. Sport has lower performance than expected, possibly because different

Table 1. Population-Level Results: AUROC score, F1-score, Accuracy and standard deviation for balanced and imbalanced datasets.

Activity	Balanced dataset						Imbalanced dataset					
	AUROC		F1-score		Accuracy (%)		AUROC		F1-score		Accuracy (%)	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Sleeping	0.62	0.04	0.56	0.20	58.2	2.1	0.62	0.05	0.63	0.14	92.2	4.2
Eating	0.51	0.03	0.79	0.13	50.9	2.1	0.51	0.03	0.01	0.01	88.8	2.3
Studying	0.55	0.03	0.59	0.21	53.0	2.2	0.57	0.04	0.11	0.09	74.3	4.5
Attending lecture	0.52	0.03	0.68	0.09	51.6	2.3	0.51	0.00	0.01	0.01	89.4	1.9
Online communication	0.56	0.03	0.64	0.21	55.1	2.5	0.57	0.02	0.03	0.02	84.7	1.8
Watching videos/TV	0.54	0.04	0.51	0.19	53.0	2.5	0.56	0.03	0.04	0.03	85.3	2.2
Sport	0.52	0.07	0.59	0.23	50.9	7.4	0.52	0.06	0.00	0.04	97.7	1.0
Shopping	0.57	0.06	0.58	0.23	55.3	5.2	0.48	0.05	0.00	0.00	97.3	0.5

Table 2. Hybrid Results: AUROC score, F1-score, Accuracy and standard deviation for balanced and imbalanced datasets.

Activity	Balanced dataset						Imbalanced dataset					
	AUROC		F1-score		Accuracy (%)		AUROC		F1-score		Accuracy (%)	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Sleeping	0.76	0.01	0.57	0.06	69.6	0.9	0.72	0.03	0.61	0.06	66.4	2.8
Eating	0.57	0.04	0.59	0.30	55.3	2.6	0.56	0.04	0.71	0.26	54.3	3.9
Studying	0.66	0.013	0.68	0.05	61.5	0.9	0.67	0.00	0.68	0.06	61.5	0.9
Attending lecture	0.62	0.04	0.69	0.22	58.5	3.9	0.60	0.04	0.70	0.22	57.8	3.1
Online communication	0.60	0.04	0.73	0.11	57.4	2.7	0.61	0.01	0.71	0.03	58.1	1.3
Watching videos/TV	0.56	0.05	0.72	0.22	53.2	4.4	0.55	0.05	0.58	0.07	54.2	4.4
Sport	0.58	0.04	0.55	0.23	55.6	3.5	0.59	0.03	0.72	0.01	56.5	3.6
Shopping	0.61	0.06	0.54	0.15	56.8	4.3	0.63	0.04	0.57	0.07	59.4	0.1

users have different accelerometer data when engaging in sports. It could also be due to home training (in the COVID-19 context) or the users not keeping their phones in their pockets while training. Therefore, the high activity levels would not be collected. The metrics could be higher after a personalized training period. Online communication has surprisingly high metrics since it regroups four different activities. One would have expected it to yield lower metrics, but as all activities are phone-related, it makes sense.

Imbalanced Dataset. Here, the accuracy is not representative of the model's performance. The F1 score is very low for all activities except sleeping. The explanation is that given the class imbalance, the model always predicts the negative class and leads to a high accuracy and a low F1 score. Shopping and sport represent no more than 5% of the dataset each, so if the model only predicts the negative class, it will lead to an accuracy of more than 95% each, which is the case and explains the low F1 score. Only sleeping yields high metrics because it is the biggest class, so the model had more data to train. Niche events are not well recognized using the population-level approach. For shopping, the AUROC score is lower than 0.5, meaning that the model is inverting the classes in some cases.

4.2 Hybrid Model

The metrics for the hybrid model can be seen on Table 2.

Balanced Dataset. 'Sleeping' has the best results, with an AUROC score of 0.76 for the reasons mentioned above. Also, the smartphone's activity when a person is sleeping is easy to recognize: the smartphone is probably placed on the bedside table for the night. Watching videos/TV and Online communication obtain a high F1 score. "Eating" 's metrics are low, which can be explained by the fact that people eating with their smartphones can behave differently (e.g., putting it away when eating with people, watching something, chatting, etc.). One would expect good results for "sport", but the AUROC score is only 0.58. The low metrics for Watching videos/TV can be explained by the difference between watching videos and TV. One can switch on the TV in the background and do something else on their phone (resulting in a different accelerometer activity), whereas watching a video on their phone means the attention is more focused on the phone, and the resulting accelerometer data can be the one of a phone standing still.

Imbalanced Dataset. As mentioned in Sect. 3.3, the training was done on an imbalanced dataset to see if niche events could be recognized. The F1 score is generally higher for this case than for the balanced dataset. However, the AUROC score is similar on the balanced and imbalanced data.

5 Discussion

In the original dataset, there were 34 different activities. However, most of them were dropped because of a lack of data from the original dataset while keeping eight informative and representative activities of the life of a student. These activities also have an impact on health and can help understand the life of a student and ultimately support their well-being via applications. Using only the accelerometer data for human activity recognition is challenging because of the complexity of the activities and the COVID situation at the data collection time: the activities are more likely to result in similar accelerometer data. The resulting AUROC and F1 scores are reasonable, given the challenge. It was noticed that a hybrid model performs better than a population-level one and that niche events are poorly recognized (imbalanced dataset case). While the results are relatively low for binary classification, the settings must be kept in mind: the data was collected in five countries, which induces a mix of people, sensor quality, and ways of using a phone. This multi-country approach should generalize well and calls for additional studies for multi-country data. Further, studies could evaluate the data quality per country and sensor to obtain more details. The original dataset was collected in real-life conditions, which also impacts the quality of data and the performance of models, but represents real conditions in which the model would be used. As mentioned, every smartphone has different components (especially in different countries/continents), and using only the accelerometer is relevant because this sensor is cheap, and most smartphones contain built-in accelerometers. Using multiple sensors was not the focus of this paper as the goal was to use the accelerometer data only. Moreover, there is not enough gyroscope or magnetometer data in the original dataset to train a model. Also, using only one sensor would spare battery life and would have a minimal impact on a smartphone's performance in the case of a health monitoring application.

6 Conclusion

In this work, raw accelerometer data from smartphones were fed into Deep Learning models to infer complex daily activities. Binary classification led to reasonable AUROC scores in the range of 0.51–0.62 with population-level models (non-personalized) and 0.56–0.76 with hybrid models (partially personalized) for eight complex activities. This work shows that it is possible to infer complex activities using only the smartphone's accelerometer and can be a baseline for a multi-country approach of Human Activity Recognition for the well-being of young adults.

Acknowledgement. This work was funded by the European Union's Horizon 2020 WeNet project, under grant agreement 823783. We also thank Darshana Rathnayake (Singapore Management University, Singapore) for discussions.

References

1. Final model of diversity. https://www.internetofus.eu/wp-content/uploads/sites/38/2021/03/D1.3-Final-model-of-diversity.pdf
2. Ann, O.C., Theng, L.B.: Human activity recognition: a review. In: 2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2014), pp. 389–393 (2014). https://doi.org/10.1109/ICCSCE.2014.7072750
3. Arif, M., Bilal, M., Kattan, A., Ahamed, S.I.: Better physical activity classification using smartphone acceleration sensor. J. Med. Syst. **38**(9), 1–10 (2014). https://doi.org/10.1007/s10916-014-0095-0
4. Bae, S., et al.: Detecting drinking episodes in young adults using smartphone-based sensors. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(2), 1–36 (2017)
5. Coorevits, L., Coenen, T.: The rise and fall of wearable fitness trackers, August 2016
6. Giunchiglia, F., et al.: A worldwide diversity pilot on daily routines and social practices (2020). University of Trento Technical report. No. #DISI-2001-DS-0 (23), 36–44, April 2021
7. Guvensan, M.A., Dusun, B., Can, B., Turkmen, H.I.: A novel segment-based approach for improving classification performance of transport mode detection. Sensors **18**(1) (2018). https://doi.org/10.3390/s18010087, https://www.mdpi.com/1424-8220/18/1/87
8. Hassan, M.M., Uddin, M.Z., Mohamed, A., Almogren, A.: A robust human activity recognition system using smartphone sensors and deep learning. Futur. Gener. Comput. Syst. **81**, 307–313 (2018). https://doi.org/10.1016/j.future.2017.11.029, https://www.sciencedirect.com/science/article/pii/S0167739X17317351

9. Jobanputra, C., Bavishi, J., Doshi, N.: Human activity recognition: a survey. Procedia Comput. Sci. **155**, 698–703 (2019). https://doi.org/10.1016/j.procs.2019.08.100, https://www.sciencedirect.com/science/article/pii/S1877050919310166, the 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2019), The 14th International Conference on Future Networks and Communications (FNC-2019), The 9th International Conference on Sustainable Energy Information Technology
10. Laput, G., Harrison, C.: Sensing Fine-Grained Hand Activity with Smartwatches, pp. 1–13. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300568
11. Likamwa, R., Liu, Y., Lane, N., Zhong, L.: Moodscope: building a mood sensor from smartphone usage patterns, June 2013. https://doi.org/10.1145/2462456.2464449
12. Mea, V.D., Quattrin, O., Parpinel, M.: A feasibility study on smartphone accelerometer-based recognition of household activities and influence of smartphone position. Inform. Health Soc. Care **42**(4), 321–334 (2017). https://doi.org/10.1080/17538157.2016.1255214, pMID: 28005434
13. Meegahapola, L., Bangamuarachchi, W., Chamantha, A., Ruiz-Correa, S., Perera, I., Gatica-Perez, D.: Sensing eating events in context: a smartphone-only approach. IEEE Access **10**(ARTICLE) (2022)
14. Meegahapola, L., Gatica-Perez, D.: Smartphone sensing for the well-being of young adults: a review. IEEE Access **9**, 3374–3399 (2020). https://doi.org/10.1109/ACCESS.2020.3045935
15. Mohamed, R., Zainudin, M.N.S., Perumal, T., Mustapha, N.: Multi-label classification for physical activity recognition from various accelerometer sensor positions (2018)
16. Plötz, T., Guan, Y.: Deep learning for human activity recognition in mobile computing. Computer **51**(5), 50–59 (2018). https://doi.org/10.1109/MC.2018.2381112
17. Ranasinghe, S., Machot, F.A., Mayr, H.C.: A review on applications of activity recognition systems with regard to performance and evaluation. Int. J. Distrib. Sens. Netw. **12**(8), 1550147716665520 (2016). https://doi.org/10.1177/1550147716665520
18. Servia-Rodríguez, S., Rachuri, K.K., Mascolo, C., Rentfrow, P.J., Lathia, N., Sandstrom, G.M.: Mobile sensing at the service of mental well-being: a large-scale longitudinal study. In: Proceedings of the 26th International Conference on World Wide Web, pp. 103–112 (2017)
19. Shoaib, M., Bosch, S., Incel, O.D., Scholten, H., Havinga, P.J.M.: Complex human activity recognition using smartphone and wrist-worn motion sensors. Sensors **16**(4) (2016). https://doi.org/10.3390/s16040426, https://www.mdpi.com/1424-8220/16/4/426
20. Straczkiewicz, M., James, P., Onnela, J.P.: A systematic review of smartphone-based human activity recognition for health research (2021)
21. Straczkiewicz, M., Glynn, N., Harezlak, J.: On placement, location and orientation of wrist-worn tri-axial accelerometers during free-living measurements, May 2019. https://doi.org/10.3390/s19092095
22. Wu, W., Dasgupta, S., Ramirez, E.E., Peterson, C., Norman, G.J.: Classification accuracies of physical activities using smartphone motion sensors. J. Med. Internet Res. **14**(5), e130 (2012). https://doi.org/10.2196/jmir.2208, http://www.jmir.org/2012/5/e130/
23. Zhu, C., Sheng, W.: Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living. IEEE Trans. Syst. Man Cybern. Part A Syst. Humans **41**(3), 569–573 (2011). https://doi.org/10.1109/TSMCA.2010.2093883

Research on Passive Assessment of Parkinson's Disease Utilising Speech Biomarkers

Daniel Kovac[1(✉)], Jiri Mekyska[1], Lubos Brabenec[2],
Milena Kostalova[2,3], and Irena Rektorova[2,4]

[1] Department of Telecommunications, Brno University of Technology,
Brno, Czech Republic
`xkovac41@vut.cz`
[2] Applied Neuroscience Research Group, Central European Institute of Technology – CEITEC, Masaryk University, Brno, Czech Republic
[3] Department of Neurology, Faculty Hospital and Masaryk University,
Brno, Czech Republic
[4] First Department of Neurology, Faculty of Medicine and St. Anne's University Hospital, Masaryk University, Brno, Czech Republic

Abstract. Speech disorders, collectively referred to as hypokinetic dysarthria (HD), are early biomarkers of Parkinson's disease (PD). To assess all dimensions of HD, patients could perform several speech tasks using a smartphone outside a clinic. This paper aims to adapt the parametrization process to running speech so that a patient is not required to interact actively with the device, and features can be extracted directly from phone calls. The method utilizes a voice activity detector followed by a voicing detection. The algorithm was tested on a database of 126 recordings (86 patients with PD and 40 healthy controls) of monologue mixed with noise with different signal-to-noise ratios (SNR) to simulate the real environment conditions. Pearson correlation coefficients show a strong linear relationship between speech features and patients' scores assessing HD and other motor/non-motor symptoms – p-value < 0.01 for the normalized amplitude quotient (NAQ) with Test 3F Dysarthric Profile (DX index) and Unified Parkinson's Disease Rating Scale (part III) in 20 dB SNR conditions, p-value < 0.01 for the jitter and shimmer with the Mini Mental State Exam (10 dB SNR). A model based on the Extreme Gradient Boosting algorithm predicts the DX index with a 10.83% estimated error rate (EER) and the Addenbrooke's Cognitive Examination-Revise (ACE-R) score with 13.38% EER. The introduced algorithm can potentially be used in mHealth applications for passive monitoring and assessment of PD patients.

This study was supported by the Quality Internal Grants of BUT (project KInG, reg. no. CZ.02.2.69/0.0/0.0/19_073/0016948; financed from OP VVV), by EU - Next Generation EU (project no. LX22NPO5107 (MEYS)), and by the Czech Ministry of Health (grant no. NU22J-04-00074).

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 259–273, 2023.
https://doi.org/10.1007/978-3-031-34586-9_18

Keywords: Hypokinetic dysarthria · Parkinson's disease · Passive assessment · Running speech

1 Introduction

Even though Parkinson's disease (PD), the chronic neurodegenerative disorder [17], was described more than two hundred years ago [22], we still do not know its exact causes. Besides genetic predisposition and age, pesticide exposure, high caloric intake, or head injuries may increase the risk of its development [6]. In addition, we are still unable to cure it, but its alleviation is possible by using medication (e.g. levodopa) [7] or other, more invasive ways (such as deep brain stimulation [3]). Moreover, there are signs that the incidence is increasing over time [26]. Thus, early diagnosis has a crucial impact on the future course of the disease, and it is essential to have the means to diagnose and monitor it at its earliest possible stage, as it can improve the patient's life.

The first symptoms of PD include speech and voice disorders [23], referred to as hypokinetic dysarthria (HD) [11], manifested by deteriorated respiration, phonation, articulation, and prosody [24]. PD patients may have impaired speech in all or only some domains [25], but at least one speech domain is affected in up to 90% of cases [14]. Tremor [30], hoarseness [5], audible breath [28], hypernasality [15], speech disfluency [18], or inappropriate silences [13] are common symptoms of HD. The speech is also often quiet [1] and unintelligible [29] and can be followed by monopitch and monoloudness [8].

Ergo, acoustic speech and voice analysis is considered an objective, supportive but practical tool for PD detection and monitoring. The analysis is done by recording the patient's speech signal with its consequent digital processing. After obtaining acoustic features that quantify individual speech disorders, it is possible to compare them with those of healthy controls (HC) using statistics to assess the severity of HD. In order to examine all domains of HD, patients are asked to perform several speech tasks, which are most commonly a sustained phonation of the vowel [a], monologue, reading text, picture description and diadochokinetic task (DDK) consisting of repeating the syllables [pa]-[ta]-[ka].

Due to the clinical time pressure, number of patients, and the inability of some patients to travel for speech recording, telemedicine plays an important role here. Smartphone applications could allow neurologists to monitor speech impairment and PD progress remotely. Patients can then perform speech tasks from home, or it could even be possible to process the speech recorded during a phone call, which would have numerous advantages.

1.1 State of the Art

Zhan et al. (2018) sought to answer whether a smartphone can be used to quantify the severity of motor symptoms of PD. To do this, they used the HopkinsPD mobile app, on which patients performed a sustained phonation of the vowel [a], from which the signal amplitude was measured. In addition to this task, patients performed four others (finger tapping, walking, balance test and reaction time test). The results correlated with the current standard rating scales [32].

Horin et al. (2019) investigated the usability of smartphone apps to treat gait, speech, and dexterity in people with PD. Regarding the HD, they measured the maximum phonation time and the mean fundamental frequency from a sustained vowel [a] task and the maximum reading time and standard deviation of the number of semitones from the fundamental frequency from a text reading task. The tasks are part of the Beats Medical Parkinsons Treatment App mobile app. Statistical tests showed no significant correlation between these features and patients' UPDRS (Unified Parkinson's Disease Rating Scale) scores or the time over which they performed the tasks [16].

Orozco-Arroyave et al. (2020) presented the Apkinson mobile app that assesses and monitors the motor skills of PD patients in terms of speech articulation, gait regularity and rigidity, and finger tapping accuracy. Speech features were extracted from the following tasks: sustained phonation of the vowels [a], [i] and [u] (jitter), DDK (articulation rate as the number of voiced segments per time and the probability which phonemes belong to each phonemic group), text reading (error between the word read and the word recognized by the automatic speech recognition model) and picture description (standard deviation of the fundamental frequency). According to the Kruskal-Walis test, the jitter and articulation rate features showed significant differences between HC and PD subjects [21].

Arora et al. (2021) analyzed 4242 smartphone recordings of a sustained phonation of the vowel [a] collected in a clinic and at home from 92 HC, 112 patients with rapid eye movement sleep behavior disorder (iRBD), and 335 patients with PD. They used acoustic signal analysis (337 phonatory features) and machine learning. Using the leave-one-subject-out cross-validation method, they could distinguish PD patients from HC with sensitivity (SEN) of 59% and specificity (SPE) of 59% [2].

Laganas et al. (2021) trained machine learning models on Mel Frequency Cepstral Coefficients and Bark-band Energies of HC and PD patients extracted from passive smartphone phone call recordings using the iPrognosis mobile application. Gender and age were added to the feature matrix, and groups of testing data were balanced in terms of these confounding factors. After leave-one-out cross-validation, the best-performing models provided an area under the curve (AUC) with the threshold operating characteristic (ROC) of 0.69/0.68/0.63/0.83 for English/Greek/German/Portuguese speaking subjects [20].

Simek and Rusz (2021) tested the effect of speech task and ambient noise (10 dB and 20 dB signal-to-noise ratio) on sensitivity to capture dysphonia of PD and iRBD patients using unsmoothed (CPP) and smoothed (CPPs) cepstral peak prominence features. There was a significant difference between PD patients and HC in sustained phonation of vowel [a] via the CPP ($p < 0.05$) and CPPS ($p < 0.01$) and the monologue via the CPP ($p < 0.01$) according to a one-way analysis of variance. The differences dropped with the addition of noise [27].

1.2 Objectives

The iPrognosis app is the only one known to have been used to detect PD from running speech recorded during phone calls. However, the employed speech features do not quantify disorders in all domains of HD. It is clear that a new approach to parametrization is needed, as existing extraction is dependent on the speech task performed. The main aim of this paper is to explore a new approach to passive HD assessment based on acoustic analysis of running speech in a noisy environment. A new method of feature extraction will be proposed so that the patient is not required to interact actively with the device and perform speech tasks. Subsequently, the robustness of features to noise that may be present in recordings during a phone call will be determined.

2 Materials and Methods

2.1 Dataset

The test database (PARCZ [12]) contains a total of 126 recordings of Czech speech (40 HC and 86 patients with PD) recorded with a condenser microphone in a regular, non-soundproofed room. Table 1 describes the representation of males and females, and Fig. 1 shows the age distribution of the subjects. During the monologue recording the participants mainly talked about their hobbies, interests, family, or profession. The mean recording length of HC is 26.3 ± 13.5 s, and for people with PD, it is 28.0 ± 15.0 s. Recordings were downsampled from 44.1 kHz to 16 kHz sampling frequency.

Table 1. Demograpfhic data.

	women	men	total
HC	21	19	40
PD	37	49	86
total	58	68	126

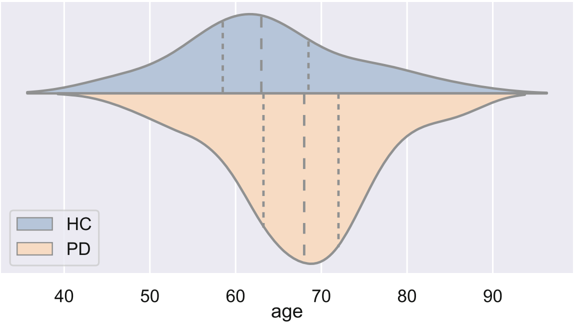


Fig. 1. Age distribution.

The participants also underwent clinical tests or questionnaires examining their motor and non-motor symptoms, sleep disorders, level of dysarthria, intelligence, depression and cognitive abilities. Results, along with the duration of the disease and medication, are shown in Table 2.

Table 2. Clinical data (mean ± std).

	HC	PD
PD duration [year]	-	14.1 ± 2.8
LED [mg]	-	987.1 ± 525.7
UPDRS III	-	24.6 ± 12.0
UPDRS IV	-	3.2 ± 2.7
FOG	-	7.1 ± 6.0
NMSS	-	39.3 ± 23.5
RBDSQ	-	3.8 ± 3.3
faciokinesis	27.9 ± 1.9	24.4 ± 3.5
phonorespiration	28.6 ± 1.4	23.8 ± 3.7
phonetics	29.5 ± 1.0	25.6 ± 3.7
overall DX index	86.0 ± 3.2	73.8 ± 9.3
IQ	-	106.7 ± 13.6
BDI	-	9.6 ± 5.9
ACE-R	-	86.4 ± 9.2
ACE-R (attention and orientation)	-	17.2 ± 1.3
ACE-R (memory)	-	19.5 ± 4.3
ACE-R (fluency)	-	10.4 ± 2.7
ACE-R (language)	-	24.9 ± 1.4
ACE-R (visuospatial)	-	14.9 ± 1.4
MMSE	28.3 ± 1.5	28.0 ± 2.5

[1] LED – Levodopa Equivalent Dose, UPDRS – Unified Parkinson's Disease Rating Scale (part III: Motor examination, part IV: Motor complications), FOG – Freezing of Gait, NMSS – Non-Motor Symptoms Scale, RBDSQ – REM Sleep Behavior Disorder Screening Questionnaire, DX index – Test 3F Dysarthric Profile (dysarthric index composed of faciokinesis, phonorespiration and phonetics), IQ – Intelligence Quotient, BDI – Beck Depression Inventory, ACE-R – Addenbrooke's Cognitive Examination-Revised, MMSE – Mini Mental State Exam

2.2 Simulation of Real Environment Conditions

The original recordings were mixed with three different types of ambient noise obtained from Freesound [10]:

☐ **Car**: the interior of a car during driving through a small city, the sound of an engine, rain and windscreen wipers.

☐ **Town square**: Union Square in San Francisco, a small art show, people chatting and moving with different objects, sometimes cars and birds.
☐ **TV**: ambient noise of an Indian television, channel changing, advertisements and music.

in 10 dB and 20 dB signal-to-noise ratio (SNR) in the following steps:

1. Resample the noise signal to the uniform 16 kHz sampling frequency.
2. Cut the noise signal to have an equal length as a speech recording.
3. Normalize both signals to have a maximum value equal to 1:

$$\mathbf{s} = \frac{\mathbf{s}}{\max(\mathbf{s})}, \tag{1}$$

   where $\mathbf{s}$ stands for the noise or the speech signal.
4. Get the power P of both signals:

$$P = \frac{\sum\limits_{n=0}^{N-1} s[n]^2}{N}, \tag{2}$$

   $s[n]$ stands for the n-th sample in a sampled audio signal of length N.
5. Get the signal-to-noise ratio SNR in dB:

$$SNR = 10 \cdot \log_{10} \frac{P_{\text{speech}}}{P_{\text{noise}}}, \tag{3}$$

   where P_{speech} and P_{noise} is the power of the speech and noise signal, respectively.
6. Mix the normalized speech signal $\mathbf{s}_{\text{speech}}$ with the normalized noise signal $\mathbf{s}_{\text{noise}}$ attenuated by the coefficient C:

$$C = \sqrt{10^{\frac{(SNR - SNR_{\text{M}})}{10}}}, \tag{4}$$

$$\mathbf{s}_{\text{mix}} = \mathbf{s}_{\text{speech}} + C \cdot \mathbf{s}_{\text{noise}}, \tag{5}$$

   where, SNR_{M} is 10 or 20 (dB) and $\mathbf{s}_{\text{mix}}$ is the final mixed signal.

Figure 2 shows a clean (original) recording of the Czech word “cestování”, meaning “travelling” in English, and the same signal mixed with noise using different SNRs.

Recordings were mixed with noise in a way so that a random third was mixed with car noise, a second third with square noise, and the last third with TV noise. That resulted in 3 datasets – a dataset of clean recordings, then noisy recordings with 20 dB SNR and finally with 10 dB SNR.

2.3 Feature Extraction

The parameterization algorithm is programmed in MATLAB environment. All features along with their description and the speech disorders they quantify are

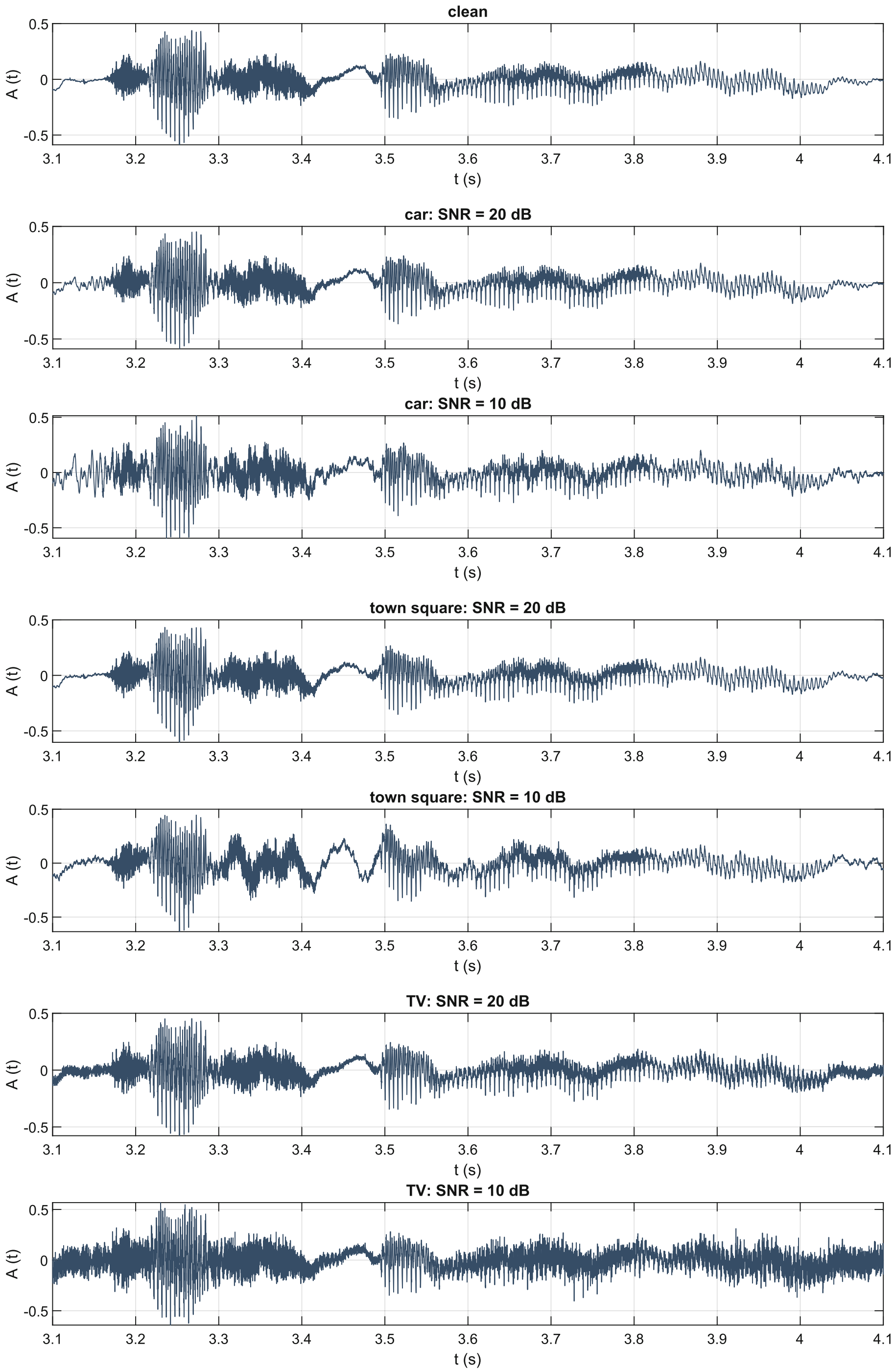


Fig. 2. A part of the clean and noisy monologue recording with different SNRs and types of noise.

Table 3. Speech features.

Acoustic feature	Specific disorder	Expected change	Feature definition
PHONATION			
CPP	Increased hoarseness	↓	Cepstral peak prominence representing the dysphonia. CPP is defined as the difference between the cepstral peak representing the fundamental frequency and the linear regression line calculated from the magnitude-quefrency cepstra.
HRF	Increased breathness	↓	Harmonic richness factor, the amount of noise in the speech signal, mainly due to incomplete vocal fold closure. HRF is defined as the ratio between the sum of magnitudes of higher order harmonics and magnitude of the fundamental frequency.
NAQ	Increased voice harshness	↑	Normalised amplitude quotient, defined as A/(D*T0), where A is the amplitude of the glotal flow pulse, D is the negative peak amplitude of the glotal flow derivative and T0 is one period of glotal flow.
relNAQSD	Rigidity of vocal folds	↑	The standard deviation of normalised amplitude quotient relative to its mean.
QOQ	Increased voice harshness	↑	Mean quasi-open quotient, defined as the ratio between the time of opened phase and fundamental period (once cycle of the vocal fold).
relQOQSD	Rigidity of vocal folds	↑	The standard deviation of quasi-open quotient relative to its mean.
jitter	Microperturbations in frequency	↑	Frequency perturbation, extent of variation of the voice range. jitter is defined as the variability of F0 of speech from one cycle to the next.
shimmer	Microperturbations in amplitude	↑	Amplitude perturbation representing rough speech. shimmer is defined as the sequence of maximum extent of the signal amplitude within each vocal cycle.
ARTICTULATION			
RFA1	Articulatory decay	↓	Resonant frequency attenuation defined as the distance in LPC spectrum between resonance of second formant and the local minima before this formant (in dB).
RFA2	Articulatory decay	↓	Resonant frequency attenuation defined as the distance in LPC spectrum between resonance of second formant and the local minima after this formant (in dB).
#loc_max	Articulatory decay	↓	The number of local maxima in transfer function of the vocal tract representing the resonances.
relF1SD	Rigidity of tongue and jaw	↓	Standard deviation of first formant relative to its mean.
relF2SD	Rigidity of tongue and jaw	↓	Standard deviation of second formant relative to its mean.
PROSODY			
RSV	Reduced number of sustained vowels	↓	The ratio of vowels sustained for longer than 100 ms to the total number of vowels (grouped voiced segments).
relF0SD	Monopitch	↓	Pitch variation, defined as a standard deviation of F0 contour of voiced segments longer than 100 ms relative to its mean.
relSE0SD	Monoloudness	↓	Speech loudness variation, defined as a standard deviation of energy of voiced segments longer than 100 ms relative to its mean.
SPIR	Inappropriate silences	↓	Number of pauses (longer than 50 ms and shorter than 2 s) relative to total speech time.
DurMED	Longer duration of silences	↑	Median duration of silences longer than 50 ms and shorter than 2 s.
DurMAD	Higher variability of silence duration	↑	Median absolute deviation of silence duration (longer than 50 ms and shorter than 2s)

listed in Table 3. It also describes the expected change in the feature for patients with increasing severity of HD. We used a Troparion [31] toolbox for extracting jitter and shimmer and Covarep [9] for the rest of the phonatory features.

The feature extraction is modified to extract the phonatory features directly from the monologue in order to examine all domains of HD together with articulatory and prosodic ones. First, a voice-activity-detector (VAD) is applied to the speech signal, followed by a voicing check. In both cases, the PRAAT [4] tool is used. If a voiced segment is longer than 100 ms, the phonatory features are extracted. Otherwise, the voiced signal is not long enough to obtain the features such as jitter or shimmer, as not enough vocal tract cycles are repeated. The fundamental frequency for someone may be less than 100 Hz, corresponding to 10 ms, and at least five cycles are needed to obtain these features. Features measuring the pausing (SPIR, DurMED, DurMAD) neglect pauses shorter than 50 ms (articulatory pauses), but also neglect pauses longer than 2 s. This length is set based on the dissertation thesis written by Tyler S Kendall [19], which describes the variation in speech rate and silent pause duration by North American English speakers. Of the 22,000 measured pauses, only some outliers exceeded the pause length of 2 s, which probably include hesitation pauses and pauses that are purely for breathing. A speech signal fragment in Fig. 3 describes the particular parametrization process.

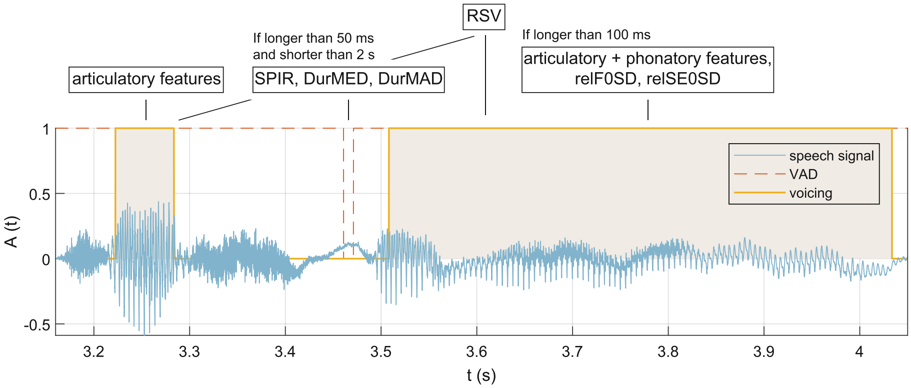


Fig. 3. Parametrization of the czech word "cestování". A zero value means that there is no speech according to the Voice Activity Detector (VAD) and/or no voiced segment.

Finally, to suppress the effect of confounding factors such as age and gender, we regressed them out. This was done using the Python programming language, which was also used for further statistical analysis and machine learning.

2.4 Statistical Analysis

To analyze the statistical relationship between speech features and clinical scores of PD patients, we calculated Pearson's correlation coefficients. The Benjamini/Hochberg test controlled the false discovery rate (FDR).

2.5 Machine Learning

By using the Extreme Gradient Boosting (XGBoost) algorithm, we mathematically modelled the extracted features during a prediction of clinical scores, or stratification the participants into the PD/HC groups. Hyperparameter tuning was included in the pipeline using a random search approach, and the models were validated using the stratified 10-fold cross-validation technique with 20 repetitions. The model's performance was evaluated by the area under the curve (AUC) of the receiver operating characteristics (ROC), sensitivity (SEN), and specificity (SPE) for the classification, and by the mean absolute error (MAE) and estimated error rate (EER) for the regression:

$$\mathrm{EER} = \frac{\mathrm{MAE}}{R}, \tag{6}$$

where R represents the range of values (of a clinical scale) in the training set. Finally, each feature's importance in predicting each clinical score was obtained to measure how valuable the feature was in building the boosting decision tree. The importance coefficients of models trained on features of each dataset (clean, 20 dB, 10 dB) were multiplied to obtain global feature importances.

3 Results

Results of the correlation analysis can be found in Table 4. It focuses only on the clinical scores, which strongly correlate with at least one speech feature in any dataset (clean, 20 dB, 10 dB). Table 5 shows the regression model's performance in predicting the clinical scores. The three most important speech features in predicting each score are also listed. The classification results are illustrated in Fig. 4 (the ROC curve represents the model's performance trained on clean or noisy data with a signal-to-noise ratio of 20 dB and 10 dB; the curves and values shown are the averages of the results from the stratified cross-validation).

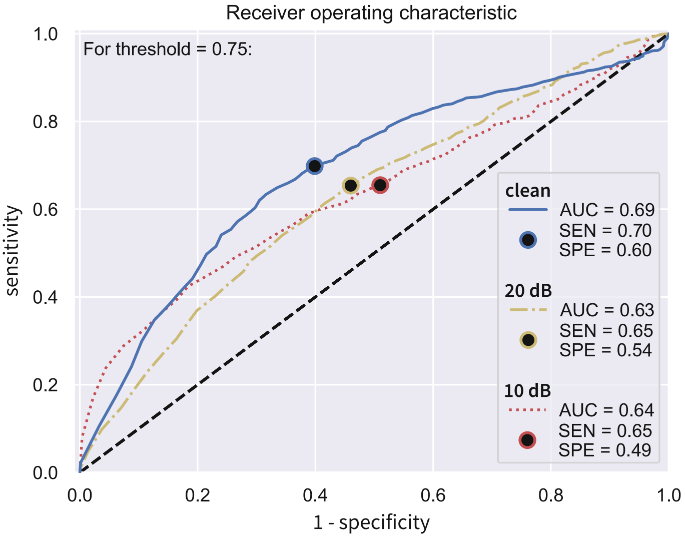


Fig. 4. Avarage ROC curve for the different signal-to-noise ratio scenarios.

Table 4. Coefficients and p-values after the FDR correction of Pearson's linear correlation between speech features and scores of clinical tests.

	UPDRS III		faciokinesis		phonoresp.		phonetics		DX index		MMSE	
	coeff	p-value	coeff	p-value	coeff	p-value	coeff	p-value	coeff	p-value	coeff	p-value
clean												
RSV	−0.125	0.432	0.078	0.815	0.000	0.999	0.133	0.219	0.077	0.495	−0.108	0.612
CPP	0.037	0.914	0.107	0.673	0.119	0.352	0.058	0.521	0.107	0.393	0.216	0.190
HRF	−0.258	0.152	0.188	0.171	0.138	0.289	0.213	0.045*	0.201	0.095	−0.105	0.612
NAQ	0.387	0.001**	−0.309	0.001**	−0.262	0.028*	−0.353	0.001**	−0.345	0.001**	0.171	0.336
relNAQSD	−0.038	0.914	0.057	0.829	0.075	0.596	0.148	0.176	0.105	0.393	−0.054	0.727
QOQ	0.197	0.219	−0.217	0.095	−0.192	0.122	−0.269	0.019*	−0.254	0.025*	0.020	0.883
relQOQSD	0.003	0.976	0.027	0.855	0.085	0.551	0.095	0.371	0.078	0.495	−0.059	0.727
jitter	−0.215	0.175	−0.027	0.855	−0.041	0.725	0.063	0.521	−0.002	0.978	−0.274	0.085
shimmer	−0.217	0.175	0.038	0.855	0.029	0.791	0.151	0.175	0.081	0.495	−0.275	0.085
RFA1	−0.008	0.976	0.104	0.673	0.087	0.551	0.169	0.125	0.134	0.255	−0.015	0.883
RFA2	0.023	0.933	0.231	0.085	0.283	0.019*	0.252	0.019*	0.288	0.010*	0.135	0.475
#loc_max	0.129	0.432	−0.037	0.855	−0.042	0.725	−0.105	0.326	−0.069	0.528	0.216	0.190
relF1SD	−0.114	0.469	0.003	0.969	−0.053	0.707	−0.087	0.393	−0.052	0.629	−0.145	0.456
relF2SD	−0.156	0.413	0.022	0.855	−0.058	0.707	−0.060	0.521	−0.037	0.717	−0.200	0.213
relF0SD	−0.050	0.914	0.085	0.815	0.151	0.269	0.255	0.019*	0.184	0.130	−0.054	0.727
relSE0SD	−0.032	0.914	0.055	0.829	0.202	0.122	0.111	0.322	0.140	0.253	0.083	0.727
SPIR	−0.217	0.175	0.112	0.673	0.197	0.122	0.244	0.023*	0.209	0.090	0.068	0.727
DurMED	0.126	0.432	−0.054	0.829	−0.148	0.269	−0.216	0.045*	−0.158	0.212	0.059	0.727
DurMAD	0.136	0.432	−0.051	0.829	−0.133	0.289	−0.209	0.045*	−0.149	0.230	0.046	0.743
SNR = 20 dB												
RSV	0.153	0.253	−0.127	0.423	−0.121	0.352	−0.123	0.401	−0.139	0.309	0.068	0.669
CPP	−0.022	0.943	0.145	0.339	0.210	0.120	0.160	0.204	0.194	0.114	0.279	0.044*
HRF	−0.218	0.139	0.150	0.339	0.061	0.864	0.112	0.407	0.119	0.357	−0.118	0.507
NAQ	0.400	0.001**	−0.276	0.038*	−0.242	0.066	−0.376	0.001**	−0.334	0.001**	0.204	0.177
relNAQSD	0.040	0.943	0.000	0.996	0.005	0.952	0.076	0.527	0.030	0.770	−0.061	0.669
QOQ	0.234	0.139	−0.234	0.057	−0.192	0.122	−0.302	0.010*	−0.271	0.013*	0.054	0.669
relQOQSD	0.022	0.943	0.168	0.290	0.168	0.193	0.210	0.066	0.204	0.104	−0.151	0.323
jitter	−0.175	0.197	−0.058	0.780	−0.140	0.323	−0.024	0.791	−0.085	0.594	−0.375	0.001**
shimmer	−0.236	0.139	0.084	0.669	−0.010	0.952	0.102	0.415	0.064	0.651	−0.394	0.001**
RFA1	0.194	0.192	−0.110	0.467	0.017	0.952	−0.073	0.527	−0.060	0.651	0.050	0.669
RFA2	0.031	0.943	0.245	0.057	0.294	0.019*	0.239	0.044*	0.292	0.010*	0.152	0.323
#loc_max	0.121	0.393	−0.035	0.780	−0.048	0.908	−0.113	0.407	−0.073	0.651	0.217	0.177
relF1SD	−0.223	0.139	0.075	0.698	−0.032	0.916	0.033	0.754	0.026	0.770	−0.153	0.323
relF2SD	−0.274	0.104	0.111	0.467	−0.033	0.916	0.063	0.542	0.050	0.684	−0.200	0.177
relF0SD	0.007	0.955	0.046	0.780	0.014	0.952	0.064	0.542	0.046	0.684	−0.002	0.987
relSE0SD	−0.006	0.955	0.039	0.780	0.192	0.122	0.076	0.527	0.118	0.357	0.089	0.635
SPIR	−0.189	0.192	0.011	0.957	0.045	0.908	0.101	0.415	0.059	0.651	−0.106	0.544
DurMED	0.181	0.197	−0.053	0.780	−0.120	0.352	−0.206	0.066	−0.143	0.309	0.073	0.669
DurMAD	0.172	0.197	−0.035	0.780	−0.119	0.352	−0.205	0.066	−0.136	0.309	0.052	0.669
SNR = 10 dB												
RSV	0.100	0.551	−0.139	0.326	−0.079	0.762	−0.098	0.542	−0.117	0.532	0.053	0.685
CPP	0.096	0.551	0.054	0.774	0.109	0.619	0.052	0.628	0.082	0.629	0.224	0.203
HRF	−0.176	0.252	0.059	0.774	0.023	0.890	0.030	0.737	0.042	0.765	−0.054	0.685
NAQ	0.302	0.076	−0.166	0.300	−0.174	0.269	−0.248	0.057	−0.220	0.076	0.149	0.380
relNAQSD	0.082	0.618	−0.243	0.114	−0.157	0.269	−0.204	0.139	−0.225	0.076	−0.008	0.938
QOQ	0.227	0.152	−0.176	0.300	−0.155	0.269	−0.244	0.057	−0.215	0.076	0.069	0.657
relQOQSD	0.070	0.662	−0.148	0.314	−0.076	0.762	−0.090	0.542	−0.116	0.532	−0.076	0.657
jitter	−0.193	0.206	−0.095	0.593	−0.161	0.269	−0.058	0.616	−0.120	0.532	−0.362	0.001**
shimmer	−0.271	0.076	0.028	0.874	−0.028	0.890	0.072	0.542	0.025	0.778	−0.351	0.001**
RFA1	0.214	0.152	−0.157	0.300	0.006	0.949	−0.102	0.542	−0.092	0.608	0.086	0.657
RFA2	−0.018	0.872	0.200	0.228	0.260	0.057	0.183	0.190	0.243	0.076	0.140	0.384
#loc_max	0.133	0.385	−0.048	0.774	−0.056	0.890	−0.135	0.498	−0.089	0.608	0.190	0.266
relF1SD	−0.218	0.152	0.091	0.593	−0.046	0.890	0.045	0.649	0.031	0.778	−0.169	0.337
relF2SD	−0.271	0.076	0.132	0.335	−0.032	0.890	0.078	0.542	0.063	0.653	−0.210	0.214
relF0SD	−0.030	0.872	0.046	0.774	0.077	0.762	0.072	0.542	0.074	0.653	−0.009	0.938
relSE0SD	0.021	0.872	0.016	0.907	0.187	0.269	0.079	0.542	0.109	0.537	0.070	0.657
SPIR	−0.046	0.798	0.063	0.774	0.038	0.890	0.082	0.542	0.068	0.653	0.074	0.657
DurMED	0.136	0.385	−0.025	0.874	0.047	0.890	−0.095	0.542	−0.025	0.778	0.162	0.337
DurMAD	0.154	0.329	−0.005	0.957	0.009	0.949	−0.119	0.542	−0.042	0.765	0.112	0.549

* – $p < 0.05$; ** – $p < 0.01$

Table 5. Results of PD duration and clinical scores prediction in different SNR scenarios (mean values from the stratified cross-validation).

	MAE			EER [%]			Important
	clean	20 dB	10 dB	clean	20 dB	10 dB	features
PD duration	2.37	2.36	2.45	26.30	26.27	27.23	shimmer, relF0SD, CPP
UPDRS III	10.71	9.97	9.85	20.61	19.17	18.95	RFA2, NAQ, relF2SD
UPDRS IV	2.45	2.32	2.32	24.48	23.21	23.17	#loc_max, NAQ, relSE0SD
FOG	5.26	4.93	4.83	26.31	24.64	24.18	RFA2, relSE0SD, shimmer
RBDSQ	2.64	2.59	2.69	20.27	19.94	20.73	NAQ, RSV, QOQ
faciokinesis	3.01	2.82	2.79	14.32	13.41	13.29	QOQ, HRF, SPIR
phonorespiration	3.03	3.07	2.84	14.41	14.61	13.51	jitter, relF2SD, RFA2
phonetics	2.56	2.76	2.69	14.21	15.32	14.97	QOQ, shimmer, SPIR
overall DX index	6.51	6.52	6.28	11.23	11.24	10.83	QOQ, jitter, shimmer
BDI	4.79	4.88	5.28	17.73	18.09	19.56	relF0SD, QOQ, RFA2
ACE-R	8.68	7.25	7.09	16.38	13.67	13.38	NAQ, RSV, relF0SD
ACE-R (attention and orientation)	0.99	1.08	1.02	19.82	21.63	20.44	relF1SD, relNAQSD, jitter
ACE-R (memory)	4.01	3.47	3.62	17.45	15.09	15.76	NAQ, relQOQSD, relF2SD
ACE-R (fluency)	2.64	2.53	2.53	24.02	23.00	23.04	NAQ, RSV, QOQ
ACE-R (language)	1.15	1.04	1.04	19.10	17.33	17.34	SPIR, relSE0SD, CPP
ACE-R (visuospatial)	1.32	1.17	1.21	22.07	19.55	20.22	relSE0SD, RFA2, relNAQSD
MMSE	1.84	1.74	1.81	13.11	12.40	12.96	CPP, NAQ, RFA2

4 Discussion

We tested a novel approach of acoustic speech feature extraction from a running speech on a database of 126 recordings (40 HC and 86 patients with PD). The algorithm was designed to be able to parametrize speech recordings obtained from phone calls, and, at the same time, to objectively assess the severity of HD in all speech domains using the obtained features. For testing purposes, the original recordings were mixed with a noise of a natural environment.

The results of the correlation analysis (Table 4) show a strong relationship between obtained features and scores of some clinical scales. These scores are of the motor skills test (UPDRS III), Test 3F Dysarthric Profile (DX index) and its parts, and cognitive function test (MMSE). The values of all significant correlation coefficients are consistent with the expected change in the feature at higher severity of HD described in Table 3. The analysis reveals that most features correlate with the results of the speakers' score of phonetics. Their increased breathiness due to incomplete vocal fold closure (HRF), increased voiced harshness (NAQ, QOQ), articulatory decay (RFA2), monopitch (relF0SD), inappropriate silences (SPIR), longer duration of silences (DurMED) and higher variability of silence duration (DurMAD) have a significant linear relationship with results of this test. The NAQ and RFA2 features strongly correlate with the overall dysarthric index and even with patients' clinically tested motor skills (UDPRS III). It is evident that with the increasing noise in the recordings, the correlation strength decreases. The phonatory feature NAQ, followed by RFA2 and QOQ, is the most robust in this sense. The Mini Mental State Exam (MMSE), which mainly examines a patient's orientation in time and place, concentration, and short-term memory, is the only non-motor test that strongly correlates with some speech features. These features quantify increased voice hoarseness (CPP) and mainly perturbations in frequency (jitter) and amplitude (shimmer), strongly correlated even in 10 dB SNR conditions.

The ability of the regression model to predict scores of clinical scales is expressed in Table 5. The MAE metric gives us directly the average deviation of the prediction from the true value. In this respect, the model is able to predict the PD duration from the extracted features of the original recordings with an error of 2.37 years. However, we need to consider the range of values within which we operate. The EER metric that accounts for this range points out that this error is 26.30%, which climbed to 27.30% when predicted based on features extracted from noisy recordings (10 dB SNR). The best-performing prediction in this term is the overall DX index, where the EER reaches 10.83% in the scenario with the noisiest recordings. This model's most important speech features are QOQ, jitter and shimmer. This shows that the quantification of phonatory disorders plays an essential role in predicting the severity of HD, and it also implies that the adaptation of the parametrization to running speech works well. The prediction of MMSE is the second most successful, with an EER of 12.96%. The most important feature here quantifies increased voice hoarseness (CPP). The robustness to noise of this feature is consistent with the results of the study by Simek and Rusz [27]. The Addenbrooke's Cognitive Examination-Revise score can be predicted with an EER of 13.38%, and the Unified Parkinson's Disease Rating Scale (part II) score with an EER of 18.95%. The NAQ feature is important in both cases.

The classifier reaches AUC = 0.69 with SEN = 70% and SPE = 60%, and it is evident that noise affects the accuracy of stratifying speakers into HC and PD groups (Fig. 4). These results are similar to the ones reported by Arora et al. [2] and Laganas et al. [20]. However, comparisons are not very appropriate here because each study used a different database.

5 Conclusion

In this paper, we investigated the potential of passive speech analysis to predict clinical scores that quantify the severity of PD. We showed that asking patients to record the commonly used sustained vowel [a] is not necessary, because the phonatory features can be extracted directly from specific voiced segments of the running speech. In addition, our approach enables important quantification of HD in other dimensions, such as articulation and prosody.

This paper is the first that deals with an adaptation of the established HD parametrization process for passive monitoring purposes. The suggested algorithm, tested on noisy speech recordings, can be used in mHealth applications and facilitate passive monitoring and assessment of PD.

The work could be continued with a more in-depth statistical analysis, including statistical hypothesis tests. In addition, it would be interesting to observe the number of patients deviating from the norms (given by healthy controls) in each speech feature. Nevertheless, the algorithm mainly needs to be validated in the wild.

References

1. Adams, S.G., Dykstra, A., Jenkins, M., Jog, M.: Speech-to-noise levels and conversational intelligibility in hypophonia and Parkinson's disease. J. Med. Speech-Lang. Pathol. **16**(4), 165–173 (2008)
2. Arora, S., Lo, C., Hu, M., Tsanas, A.: Smartphone speech testing for symptom assessment in rapid eye movement sleep behavior disorder and Parkinson's disease. IEEE Access **9**, 44813–44824 (2021)
3. Baudouin, R., Lechien, J.R., Carpentier, L., Gurruchaga, J.M., Lisan, Q., Hans, S.: Deep brain stimulation impact on voice and speech quality in Parkinson's disease: a systematic review. Otolaryngol. Head Neck Surg. 01945998221120189 (2022)
4. Boersma, P., Weenink, D.: PRAAT: doing phonetics by computer [computer program]. version 5.3. 51 (2013). http://www.praat.org/retrieved. Accessed 12 (2013)
5. Cernak, M., Orozco-Arroyave, J.R., Rudzicz, F., Christensen, H., Vásquez-Correa, J.C., Nöth, E.: Characterisation of voice quality of Parkinson's disease using differential phonological posterior features. Comput. Speech Lang. **46**, 196–208 (2017)
6. Chade, A., Kasten, M., Tanner, C.: Nongenetic causes of Parkinson's disease. Parkinson's Dis. Relat. Disord. **70**, 147–151 (2006)
7. Connolly, B.S., Lang, A.E.: Pharmacological treatment of Parkinson disease: a review. JAMA **311**(16), 1670–1683 (2014)
8. Darley, F.L., Aronson, A.E., Brown, J.R.: Differential diagnostic patterns of dysarthria. J. Speech Hear. Res. **12**(2), 246–269 (1969)
9. Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S.: COVAREP-A collaborative voice analysis repository for speech technologies. In: 2014 IEEE International Conference on Acoustics, Speech And Signal Processing (ICASSP), pp. 960–964. IEEE (2014)
10. Font Corbera, F., Roma Trepat, G., Serra, X.: Freesound technical demo. In: MM 2013. Proceedings of the 21st ACM International Conference on Multimedia; 21–25 Oct 2013 Barcelona, Spain. New York: ACM; 2013, p. 411–412. ACM Association for Computer Machinery (2013)
11. Freed, D.B.: Motor Speech Disorders: Diagnosis and Treatment. Plural Publishing, San Diego (2018)
12. Galaz, Z., et al.: Prosodic analysis of neutral, stress-modified and rhymed speech in patients with Parkinson's disease. Comput. Methods Programs Biomed. **127**, 301–317 (2016)
13. Hammen, V.L., Yorkston, K.M.: Speech and pause characteristics following speech rate reduction in hypokinetic dysarthria. J. Commun. Disord. **29**(6), 429–445 (1996)
14. Ho, A.K., Iansek, R., Marigliani, C., Bradshaw, J.L., Gates, S.: Speech impairment in a large sample of patients with Parkinson's disease. Behav. Neurol. **11**(3), 131–137 (1998)
15. Hoodin, R.B., Gilbert, H.R.: Nasal airflows in parkinsonian speakers. J. Commun. Disord. **22**(3), 169–180 (1989)

16. Horin, A.P., McNeely, M.E., Harrison, E.C., Myers, P.S., Sutter, E.N., Rawson, K.S., Earhart, G.M.: Usability of a daily mhealth application designed to address mobility, speech and dexterity in Parkinson's disease. Neurodegenerative Dis. Manage. **9**(2), 97–105 (2019)
17. Hornykiewicz, O.: Biochemical aspects of Parkinson's disease. Neurology **51**(2 Suppl 2), S2–S9 (1998)
18. Juste, F.S., Sassi, F.C., Costa, J.B., de Andrade, C.R.F.: Frequency of speech disruptions in Parkinson's disease and developmental stuttering: a comparison among speech tasks. PLoS ONE **13**(6), e0199054 (2018)
19. Kendall, T.S.: Speech Rate, Pause, and Linguistic Variation: An Examination Through the Sociolinguistic Archive and Analysis Project. Duke University, Durham (2009)
20. Laganas, C., et al.: Parkinson's disease detection based on running speech data from phone calls. IEEE Trans. Biomed. Eng. **69**(5), 1573–1584 (2021)
21. Orozco-Arroyave, J.R., et al.: Apkinson: the smartphone application for telemonitoring Parkinson's patients through speech, gait and hands movement. Neurodegenerative Dis. Manage. **10**(3), 137–157 (2020)
22. Parkinson, J.: An essay on the shaky palsy. London: Sherwood, Neely and Jones, pp. 1–6 (1817)
23. Poewe, W.: Global Scales to Stage Disability in PD: the Hoehn and Yahr scale. Rating Scales Parkinsons Disease, pp. 115–122 (2012)
24. Rohl, A., Gutierrez, S., Johari, K., Greenlee, J., Tjaden, K., Roberts, A.: Chapter 7 - speech dysfunction, cognition, and Parkinson's disease. In: Narayanan, N.S., Albin, R.L. (eds.) Cognition in Parkinson's Disease, Progress in Brain Research, vol. 269, pp. 153–173. Elsevier (2022). https://doi.org/10.1016/bs.pbr.2022.01.017, https://www.sciencedirect.com/science/article/pii/S0079612322000176
25. Rusz, J., Tykalova, T., Novotny, M., Ruzicka, E., Dusek, P.: Distinct patterns of speech disorder in early-onset and late-onset de-novo Parkinson's disease. npj Parkinson's Dis. **7**(1), 1–8 (2021)
26. Savica, R., Grossardt, B.R., Bower, J.H., Ahlskog, J.E., Rocca, W.A.: Time trends in the incidence of Parkinson disease. JAMA Neurol. **73**(8), 981–989 (2016)
27. Šimek, M., Rusz, J.: Validation of cepstral peak prominence in assessing early voice changes of Parkinson's disease: Effect of speaking task and ambient noise. J. Acoust. Soci. Am. **150**(6), 4522–4533 (2021)
28. Thijs, Z., Watts, C.R.: Perceptual characterization of voice quality in nonadvanced stages of Parkinson's disease. J. Voice **36**(2), 293.e11-293.e18 (2020)
29. Tjaden, K., Wilding, G.: Effects of speaking task on intelligibility in Parkinson's disease. Clin. Linguist. Phonetics **25**(2), 155–168 (2011)
30. Tykalová, T., Rusz, J., Švihlík, J., Bancone, S., Spezia, A., Pellecchia, M.T.: Speech disorder and vocal tremor in postural instability/gait difficulty and tremor dominant subtypes of Parkinson's disease. J. Neural Transm. **127**(9), 1295–1304 (2020)
31. Vashkevich, M., Petrovsky, A., Rushkevich, Y.: Bulbar ALS detection based on analysis of voice perturbation and vibrato. In: 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), pp. 267–272. IEEE (2019)
32. Zhan, A., et al.: Using smartphones and machine learning to quantify Parkinson disease severity: the mobile Parkinson disease score. JAMA Neurol. **75**(7), 876–880 (2018)

Millimeter Wave Radar Sensing Technology for Filipino Sign Language Recognition

Jorelle Aaron Herrera, Almira Astrid Muro, Philip Luis Tuason III(✉), Paul Vincent Alpano, and Jhoanna Rhodette Pedrasa

University of the Philippines Diliman, Quezon City, Philippines
`{jorelle.aaron.herrera,luis.tuason}@eee.upd.edu.ph`, `afmuro@up.edu.ph`

Abstract. Filipino Sign Language (FSL) is the primary language used by the Deaf and Hard-of-Hearing (DHH) community in the Philippines. The lack of support for FSL from the government has led to a huge communication gap between the DHH and the hearing society. A substantial amount of research has been done to develop sign language recognition systems based on computer vision or wearable technology. However, most such systems are limited to controlled settings, while wearable systems also raise issues such as inconvenience to users. Millimeter wave (mmWave) technology has recently seen potential in gesture recognition applications as it allows the system to be non-contact and resistant to environmental factors while ensuring high resolution for recognizing small movements. An mmWave-based FSL recognition system that can translate isolated signs into their equivalent gloss was developed. Data from a TI IWR1443 radar sensor was fed into a preprocessing algorithm and a deep learning model composed of multi-view 2D CNNs and LSTM. 4 models were trained based on a dataset of 24 FSL signs gathered with 3 native Deaf signers in 3 different environments. A total of 3240 samples were collected, resulting in a model that attained an overall peak accuracy of 94.9% and an average real-time recognition latency of about 2.01 s. The model's performance is comparable to both existing FSL and mmWave systems, showing immense potential for future work on FSL recognition using mmWave.

Keywords: Sign Language Recognition · Millimeter Wave · Deep Learning

1 Introduction

Filipino Sign Language (FSL) is a complex visual language composed of forming hand shapes and movements mixed with non-manual signals such as facial

---

Supported by Philippine Federation of the Deaf.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 274–288, 2023.
https://doi.org/10.1007/978-3-031-34586-9_19

expressions and upper body movements. It is the primary mode of communication used by the Deaf and Hard-of-Hearing (DHH) community in the Philippines. FSL was initially based on the American Sign Language (ASL) but has since evolved to be an independent language with the continuous addition of new local signs with each generation of the Deaf [11].

As of 2009, 1.23% of the Philippine population is either deaf, mute, or hearing-impaired, 517,536 of whom have some degree of deafness [1]. For the past decades, the Deaf community has greatly lacked needed support from the government. Only recently was FSL recognized as the national sign language of the Philippines through R.A. 11106, also known as the FSL Act of 2018, which was a watershed moment for FSL and Deaf culture in the country [22].

Living in a hearing-centric society, the DHH community is exposed every day to several environmental vulnerabilities, which include communication barriers, additional disabilities, and a lack of mental health services. Furthermore, since the COVID-19 pandemic began, their need for social support has intensified as everyone started working and living remotely–mainly relying on the internet and technological devices, the content of which caters mostly to the hearing community. Interpreting languages is crucial for them to survive, and the lack of interpreters poses a demand for technology that can facilitate a relay service for the Deaf [21].

Efforts to build interpreting technologies have been going on for the past decades as a number of FSL recognition research has been conducted in the Philippines. The most used detection medium in past studies is computer vision through cameras [4,6] and infrared sensors like Microsoft Kinect [23] while other studies have tried using wearables such as gloves equipped with sensors for more precision [17]. More recent global developments have utilized radar sensing such as Wi-Fi signals and millimeter-wave (mmWave) signals for ASL recognition and other gesture recognition applications. This type of technology holds an advantage over cameras and wearables due to their limitations such as causing discomfort of wearing gloves, having limited range, requiring ideal lighting conditions, and raising potential privacy concerns [14].

Millimeter-wave signals have been found to show more potential than Wi-Fi signals due to their ability to detect finer movements and their robustness to environmental factors such as the movement of other people and objects. [32]. Furthermore, the recent global deployment of 5G gave rise to more research on technologies that it utilizes, such as mmWave frequencies. The future expansion of 5G networks will make mmWave infrastructures more accessible to the point where they will be collaborating with today's communication infrastructures, proving its huge potential in various applications [5].

2 Related Work

2.1 Filipino Sign Language

Filipino Sign Language (FSL) is a complete, natural visual language used by a majority of the Filipino Deaf community. FSL, much like all sign languages, has its own linguistic rules for pronunciation, word formation, and order. Much like how hearing persons have different ways of speaking, signers also express themselves differently. It comes with regional variations and dialects that differ down to the smallest but most significant parameters of a sign. Other sociological factors such as geographical location, age, and gender contribute to the variety and growth of sign language [18].

The fundamental structure of a sign mainly revolves around the model developed by Liddell and Johnson in 1989, describing the 5 parameters: hand shape, location, palm orientation, movement, and non-manual signals [13]. Hand shape pertains to the arrangement of the fingers and joints, while location refers to the position of the hands relative to the body [26]. Palm orientation refers to the direction the palm is facing, and movement can refer to the movement of the fingers or the path that the hand or arms take [26]. These 4 parameters make up the manual markers of sign language. The fifth parameter is composed of the non-manual signals. Some non-manual signals that have been recorded in FSL signs include even the smallest of movements and expressions on the face, such as brow and lip movements, eye gazes, and nose wrinkling [26].

2.2 Sign Language Recognition

Sign language recognition is generally classified into two approaches: glove-based and computer vision-based (CV-based) [30]. The main advantages of gloves are their higher accuracy and hand information extraction since the sensors are directly attached to the hand. However, among its limitations are its inability to provide other essential information, such as non-manual signals and movements in the rest of the body, and the inconvenience of wearing gloves, which may restrict the movement or expression of the signer [30]. The CV-based approach, on the other hand, has been widely researched due to the ubiquity of computer vision and the rapidly rising trends in machine learning and artificial intelligence [20]. However, it still comes with its limitations, such as its sensitivity to lighting conditions [6] and other environmental factors such as unwanted objects in the video [14].

2.3 Millimeter Wave

Millimeter wave (mmWave) refers to the spectrum between 30 and 300 GHz, which has wavelengths in the millimeter range (1 to 10 mm) [3,12,16]. This frequency range has been used to develop radar sensors that could measure range, velocity, and angle by transmitting electromagnetic waves and comparing them to the received reflections of those transmitted signals [12]. Because of its high frequency, mmWave radar sensor implementations have small and closely-spaced antennas, which allow favorable characteristics such as smaller component sizes, greater availability of bandwidth, lower mutual interference between radars, and higher spatial resolution over other radar sensing technologies [3,16].

One of the many applications of mmWave radar sensing technologies is sign language recognition. This technology has been particularly on the rise in this application because of its notable advantages over camera-based and wearable device-based implementations, which include non-intrusive, device-free, and environment-resilient sensing [16,25,32]. Recent trends focus on optimizing aspects that make it viable for real-world applications, which include real-time [15,25,29,32], person-independent [15], environment-resilient recognition [25,32].

2.4 Recognition Algorithms

Millimeter wave sensors can produce different types of datasets which researchers can utilize—point clouds [14,19,25,29], continuous range doppler image sequences or spectrograms [27,32,33], and raw vibrations [7]. Among these, the most common datasets used are in the form of point clouds, which are scatter plots in the euclidean three-dimensional space.

Preprocessing Algorithms. With the mmWave sensors removing many of the points reflected by static objects through the built-in static clutter removal algorithm CFAR, the data points are from dynamic objects detected by the sensor as well as scatterings and reflections from the environment. Some of these data points can be noise in the form of outliers, which need to be removed. Most studies that used point clouds as a dataset [25] applied DBSCAN [8] or an improvement of this algorithm [14,31] for their outlier removal.

Deep Learning Architectures. Both spatial and temporal properties of the data are essential to gesture recognition applications. Consequently, learning these two properties concurrently is fundamental to the architecture to be built for this specific project. Various deep learning architectures in gesture recognition studies with millimeter wave as the medium involve neural networks, from basic convolutional neural network (CNN) models [7,32,33] to improved novel networks [14,19,29], for learning feature representations and long short-term memory (LSTM) modules [25,27,31] for modeling signs over time. Of all the architectures, CNNs with or without LSTM modules are the most common deep learning algorithms used in mmWave gesture recognition regardless of data type [7,27,31–33].

3 System Design

The overall system is composed of the radar sensor connected to a local machine containing the recognition module and graphical user interface (GUI), as illustrated in Fig. 1. The sensor is initialized through the GUI, which establishes a serial connection. Then, the system captures raw data from a user performing a sign in front of the radar sensor. The captured data is then fed into the recognition module, starting with the preprocessing algorithm to clean the data. It is then fed through the deep learning model to classify the data and return the gloss of the sign. The gloss is then displayed on the local machine through the GUI.

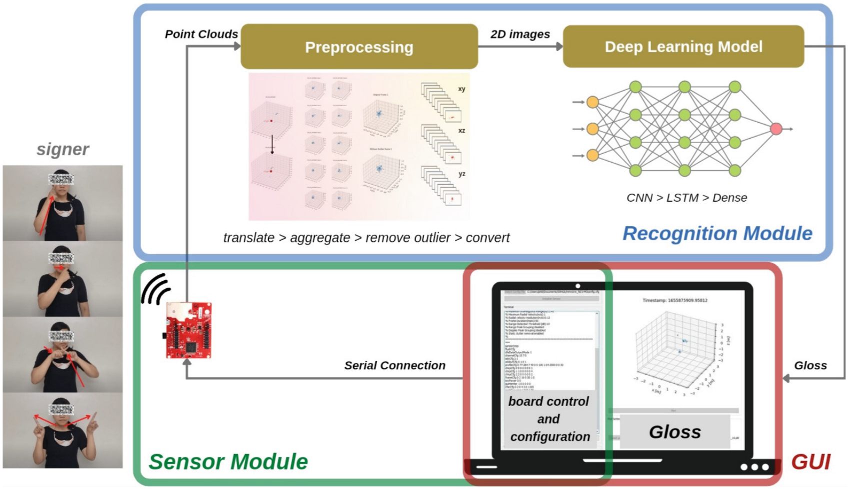


Fig. 1. Full system setup in real-time

3.1 Sensor Module

The sensor module used in this project is the IWR1443BOOST evaluation board which is a commercial off-the-shelf board from Texas Instruments that uses frequency-modulated continuous wave (FMCW) radars for high precision sensing at frequencies from 76 to 81 GHz.

Generation of Point Clouds. This board is a multiple-input and multiple-output (MIMO) device that makes use of range, elevation, and azimuth angle to generate point clouds in three-dimensional space [29]. The processes involved include Range-FFT (1D), Doppler-FFT (2D), Constant False Alarm Rate (CFAR), and Angle-FFT (3D) as shown in Fig. 2 [25,29].

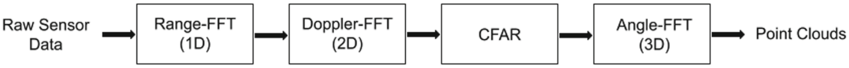


Fig. 2. Signal processing of IWR1443

Radar Configuration. The radar configuration file for the sensor module determines the characteristics of its transmitted signal and how it processes the signals it receives. The researchers configured the chirp properties to 20 fps, a velocity resolution (v_{res}) of 0.13 m/s, a range resolution (R_{res}) of 0.047 m, a maximum velocity (v_{max}) of 1.0 m/s, and a maximum range (R_{max}) of 2.41 m. In addition to this, the researchers also decided to turn off the Range Peak

Grouping and Doppler Peak Grouping, which reduce the number of data points by grouping those that are close together, to avoid the omission of necessary data points for feature extraction, and to turn on the Static Clutter Removal, which removes data points that are not in motion, to allow the system to be more resilient to noise caused by the multipath effects brought by static objects within the field-of-view of the sensor module.

3.2 Recognition Module

The recognition module is composed of two submodules, namely the preprocessing algorithm, which was inspired by the Pantomime study [25], and the deep learning model, which was derived from the ExASL study [31].

Preprocessing Module Algorithm. There are four stages in the preprocessing algorithm, as shown in Fig. 3. First, all points in all raw frames are translated to the origin. Once centered, the algorithm reduces the frames of each sample to the desired number by aggregating. Each of the aggregated frames is then subjected to noise reduction by removing outliers. Lastly, the 3D frames are converted into 2D multiview frames.

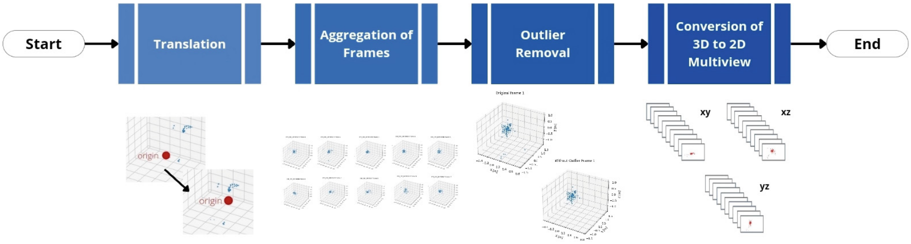


Fig. 3. Overview of the Preprocessing Algorithm

Translation. Raw point cloud data was translated into the main cluster based on the relative displacement of its centroid to the origin. At this stage, the position of the cloud is normalized, aiming to reduce the effect of the position of the signer.

Aggregation. To reduce the complexity of the module, the original number of frames of raw data from the sensor operating at 20 fps was reduced to 10 frames per sign. The method of aggregation used in this study is inspired by the time decay method of the Pantomime study [25]. All points were chronologically grouped into sets of k/f points, where k is the total number of points in the data and f is the desired number of frames per sign. In this study, the value of k varies from sign to sign while $f = 10$. Aggregating the frames resulted in a denser point cloud per frame, which enabled better outlier identification.

Noise Reduction. Point cloud data can contain noise due to scattering and reflections from the environment, especially in cluttered environments [25]. With the sensor having the function to remove points from static objects, noise from the sensor can be removed by simple outlier detection through the DBSCAN algorithm using $\epsilon = 0.5$ and minimum samples in a cluster ($min\_samp = 5$).

Conversion. The first stage of the layers of the deep learning model, CNN, required inputs of consistent and fixed dimensions. In addition, the model used 2D CNN, which takes in two-dimensional images. To meet these requirements, after aggregating the frames, the 3D point cloud data was converted into 2D multiview data. Subsequently, each sample has 3 view sets—xy-view, yz-view, and xz-view—and each view set has f frames.

Deep Learning Model. The structure of the model used in this study is derived from the study ExASL [31]. The model has three main components, which are the view-specific CNNs, LSTMs, and dense layers. The view-specific CNN portion contains 4 sets of (1) convolutional layers made of 5×5 convolutional kernels, (2) a max pooling layer with a 2×2 kernel, and (3) rectified linear units as activation arranged sequentially. Bidirectional LSTMs contain two layers of LSTM cells with 2048 hidden units. Lastly, the dense layer consists of three linear layers with 2048, 1024, and 512 hidden units, respectively. A dropout rate of $p = 0.65$ was also used for regularization [31].

3.3 Graphical User Interface

A simple graphical user interface (GUI) for the system was developed using Python 3 and PyQt5. The GUI was used in all stages of the project, namely interfacing the sensor with the local machine, facilitating the data-gathering process, and determining the feasibility of real-time applications of the system. It is able to display the console outputs of the board and displays the sensor's collected data through a 3D scatter plot. It is also used to load the deep learning model, allowing for the direct recognition of the signs.

3.4 Evaluation Parameters

Two parameters were considered for analyzing the performance of the system—sign-to-gloss accuracy (individual A_{S-G} and overall μ_A) and recognition latency (l_R). The individual sign-to-gloss accuracy, A_{S-G}, pertains to the accuracy of each of the 24 signs per model (Eq. 1), while the overall accuracy, μ_A, pertains to the accuracy of a model on all 24 signs (Eq. 2).

The recognition latency of the system, l_R, was measured by getting the mean time interval between the time when the data stream is passed from the sensor to the local machine (t_{SL}) and the time the user interface presents a final sign-to-gloss translation for the said sign (t_{UI}), with N being the total number of trials (Eq. 3).

$$A_{S-G} = \frac{\#\, of\, times\, G\, (gloss)\, is\, matched\, to\, this\, S\, (sign)}{total\, \#\, of\, attempts\, for\, S} \times 100\% \qquad (1)$$

$$\mu_A = \frac{total\ \#\ of\ correct\ translations}{total\ \#\ of\ trials} \times 100\% \qquad (2)$$

$$l_R = \frac{\sum_{n=0}^{N} t_{UI} - t_{SL}}{N} \qquad (3)$$

4 Testing

4.1 Selection of Signs and Signers

A set of 24 signs was selected by Dr. Liza Martinez, the founder and former director of the Philippine Deaf Resource Center, and approved by the Philippine Federation of the Deaf (PFD). The selection was based on the signs' distinct features with respect to some basic characteristics of the phonological structure of a sign, namely the number of articulators, the location, the path, and the movement.

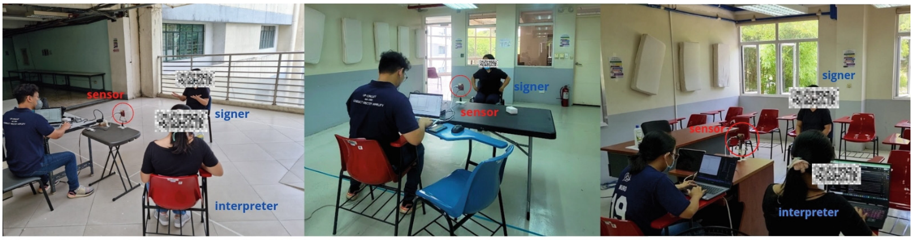
Three right-handed Deaf adults with similar signing styles volunteered, as coordinated with the PFD, as the project's participants to perform the signs in front of the system. In order to minimize potential variations in the manner of signing, the participants were selected to be native Deaf signers, meaning that they began signing in early childhood.

4.2 Physical Setups

This study defines noise as the multipath effects of static objects in the environment. To investigate the system's usability in a real-world setting and analyze the effect of the openness of an area on the signal propagation to and from the radar sensors, three different environments were implemented—(1) outdoor or open space (Fig. 4a), (2) indoor with minimal noise or an enclosed space with negligible background objects (Fig. 4b), and (3) indoor with noise or an empty classroom with static clutter (Fig. 4c). The order of pictures in Fig. 4 ranks the physical setups by noise exposure, from left having the least noise to right. In each setup, a chair was placed 1.5 m directly in front of the sensor.

4.3 Comparative Testing

Each of the three Deaf signers performed the 24 signs 15 times in each of the three environments, amounting to 1080 samples per environment and 3240 samples overall. This produced four datasets, one for each environment and one overall dataset containing all the samples from the three environments. These four datasets were used to train four individual models and were labeled as I for indoor with minimal noise, IWN for indoor with noise, O for outdoor, and C for combined.



(a) Outdoor (b) Indoor with Minimal Noise (b) Indoor with Noise

Fig. 4. Physical Setups

The four datasets were split into a ratio of 80:20 for training and testing sets, respectively. The models were trained with $batch\_size = 5$ and $epochs = 400$. An Nvidia RTX 3060 GPU with a 12 GB VRAM was used to train the models. After finalizing the models, all four of the test datasets were used to test each model's performance in terms of accuracy. The model with the highest accuracy was set to be the best model and was used in the real-time testing phase.

4.4 Real-Time Testing

The researchers conducted simple tests to determine the semi-real-time capability of the system. For logistical convenience, one of the researchers served as the signer and the setup was an outdoor environment located at the home of one of the researchers. Using the GUI and the best model achieved from comparative testing, the researchers recorded and tested 3 repetitions of each of the 24 signs. After each recording, the system automatically inferred the gloss corresponding to the performed sign with some delay. These delays were measured to compute the overall recognition latency of the system.

5 Results and Discussion

5.1 Performance of Models

Given that all test datasets were used to test each model, the performance of each model was analyzed by focusing on both the category testing results and cross testing results. This study defines *direct testing performance* as the performance of the models using the same test dataset used for training, i.e., the same environment, while *cross testing performance* is the performance of the models using datasets different from their respective categories.

Direct Testing Performance. Model O yielded the highest accuracy among the environment-separate models. This is consistent with the condition of the outdoor setup arranged in this study—an open space—which reduced the multipath effects of static objects such as walls in indoor setups. Comparing the three

Table 1. Overall Model Accuracies

Model	Accuracy (μ_A)
O	93.52%
I	88.89%
IWN	87.96%
C	**93.83%**

environment-separate models by accuracy as seen in Table 1, Model I yielded higher accuracy than Model IWN, but lower accuracy than Model O as expected based on the different levels of noise present in each setup. Therefore, these results show that the lower the noise level, the higher the model's accuracy.

Model C produced the highest accuracy (Table 1) among the four models since it was trained to classify glosses with 3 different levels of noise and 3 times the number of training samples processed than any other model. This suggests that the more data and variation used to train the model, the better it may translate signs accurately and the more resilient it can be to noise. Consequently, Model C was used in real-time testing where the researchers tested the system's real-time capability.

Ranking the environments with increasing noise levels, the outdoor setup comes first, followed by the indoor with minimal noise and the indoor with noise setup. The results in Table 2 are consistent with these setup conditions. Model I was trained using a dataset with little-to-no noise present, a noise level that can range from that of the outdoor dataset and the indoor with noise dataset. As a consequence, Model I yielded the highest accuracy and best cross testing performance among the other two models.

Table 2. Cross Testing Accuracies of Models

Model	**Test Dataset**			
	Outdoor	Indoor w/ Min. Noise	Indoor w/ Noise	Combined
O	■	56.94%	64.35%	71.60%
I	71.30%	■	72.22%	77.47%
IWN	62.96%	54.17%	■	68.36%
C	**94.91%**	**93.06%**	**93.52%**	■

Cross Testing Performance. Model O yielded the best direct testing performance among the three models that are environment-separate, but the data used to train this model contains less noise than those of other setups. Hence, it performed the worst when tested with data with noise. This phenomenon is also evident in the cross testing results of Model IWN, with the most noise. Lastly, since Model C has the most exposure to different levels of noise, it yielded the best cross testing performance among all models (Table 2).

5.2 Analysis of Signs

While the overall accuracy of the models describes the overall performance of the model, individual sign-to-gloss accuracies describe how the model behaves with respect to each sign. Analyzing this metric provides insight into how well the model can read into the phonological features of the sign that were highlighted in this project. The average individual sign-to-gloss accuracy across all the models was measured to be 90.82%, showing that each model is sufficiently able to recognize and distinguish all 24 signs from one another. For further analysis, the signs were grouped according to their individual sign-to-gloss accuracy and the common features within such groups were identified in Table 3.

Since the mmWave radar sensor collects sparse point clouds as data, the system primarily relies on location and movement for recognizing signs. Signs that have a high recognition rate in the dataset are those with more straightforward or distinct paths and sequences of movements. On the other hand, the signs with similar small movements, such as twisting of the wrist and moving of fingers, have poorer performance. Nevertheless, the limited vocabulary of the system also limits the conclusions that can be drawn based on the structure of the signs. FSL, like all sign languages, is a very complex language whose lexicon continues to evolve today. These mispredictions made by the system can be attributed to other factors such as the environment, the signers, or the size of the datasets.

Table 3. Individual Sign-to-Gloss Accuracy of Model C

Sign to Gloss Acc (A_{S-G})	Glosses	Common features between at least 2 signs within the group
100%	FEEL_LAZY, MRT_LRT, PRETEND, COVID19, CONCLUSION, PINEAPPLE, EYE_EYE_DIFFERENT	Mostly one handed Single and straight paths; Multiple movement-hold segments
96.30%	MAID, WHERE, AGREE, UTANG, RENT, I_VISIT_YOU	Same path; Similar position of dominant and non-dominant hands
92.59%	18, WALA_PERA, LOLO_LOLA, YEAR	Twisting of wrist; Similar handshape
88.89%	YES, SAME, EXPOSE, OBSERVE,	Similar position and movement of both hands
81% to 86%	CIVIL_MARRIAGE, ROOF_TWIST, COUNT	Similar finger internal movement

5.3 Feasibility of Real-Time Implementation

For the real-time testing, a total of 72 samples were recorded, 3 repetitions of each of the 24 signs, performed by one of the researchers as the signer. The average recognition latency was measured to be 2.0086 s. The best performing model, Model C, and a laptop with a 3.1 GHz Dual-Core Intel Core i5 CPU and without a GPU were used for this test. In addition, out of the 72 signs recorded, 49 were inferred correctly by the system, which resulted in an accuracy of 68.06%. The significantly lower accuracy measured in this test can be attributed to the signing experience of the signer. The dataset used to train the model was made with native Deaf signers, while a beginner signer did this test.

5.4 Comparison with Existing Studies

Table 4 shows the best accuracy of this project compared with that of relevant works in FSL recognition and mmWave systems. This project is shown to outperform much of the previous work in FSL and it is comparable with that of other mmWave systems.

Table 4. Comparison with Results of Related Work

FSL Projects		mmWave Projects	
Project and Focus	Best Accuracy	Project and Data Type	Best Accuracy
mmWave FSL (this project)	**94.91%**	[2], Spectrogram	95.04%
[6] Isolated signs	89.00%	[9], 3D Point Cloud	95.00%
[20] Static FSL numbers	83.10%	[10], Spectrogram	72.50%
[23] Basic FSL signs using arms	95.00%	[25], 3D Point Cloud	96.12%
[24] Alphabet, numbers, 30 words	79.44%	[31], 3D Point Cloud	92.50%
[28] Facial expressions	76.00%	[32], Spectrogram	86.7%

Comparison with FSL Recognition Systems. According to Cabalfin et al. [6], limiting the signs used to a small number of very distinct features results in higher accuracy. This is reflected in this project, where the signs that are distinct and straightforward tend to be recognized more accurately. In contrast, those with similarities in a number of FSL phonological parameters are less reliably recognized. In particular, this project struggles with signs with similar positions and movement, which is avoided by Oliva et al. [23] by only using signs that can be performed with the use of the arms only, and with open-palm hand shapes. This may explain why its accuracy is as high as 95% even without using deep learning models, which generally have higher accuracies compared to those that use supervised machine learning techniques like SVM [32].

Comparison with Other MmWave Recognition Systems. The results from *Pantomime* [25] show that the highest accuracy model is trained in the open area and office setups and is tested in the open area. This is reflected by this project, as the best accuracy was achieved when the model was trained on the combined dataset and tested on the outdoor dataset, which is the most open setup.

The projects, *mmASL* [32] and *ASL Recognition Using RF Sensing* [10], tested the impact of different signers on the performance of models. [10] proved that the performances of models trained and tested on native signers and non-native signers are significantly different, while [32] observed that particular samples where the participants performed the sign slower than average produced errors. These observations were also reflected in the results of this study, particularly in the real-time testing phase.

6 Conclusion and Future Work

A functional recognition system for FSL was implemented using an mmWave FMCW radar sensor and deep learning. The sensor module collected signer data in the form of point clouds, and a preprocessing algorithm converted the point clouds into multi-view 2D images. The CNN and LSTM-based deep learning models, derived from ExASL [31], were trained and tested using the collected data. A graphical user interface was used to collect data and display the output point clouds, predicted gloss, and latency of prediction.

A dataset containing 3240 total samples of 24 signs performed by three (3) native FSL signers was used to train four (4) models (I, IWN, O, C) which corresponded to the different test environments. Model C, which had the most exposure to different levels of noise and the largest number of samples in its train dataset, achieved the best overall accuracy, best direct testing performance ($\mu_A = 93.83\%$), and best cross testing performance, which ranged from 93.06% to 94.91%. Consequently, Model C was used to test the real-time capability of the system, which yielded a recognition latency of ≈ 2.01 s. The findings are consistent with that of the existing FSL and mmWave recognition systems, and the model performance is competitive when compared to the same projects, especially in FSL recognition.

Improvements can be made in the various aspects of the study for future iterations. Introducing more variations, namely the number of signs, signers, and different environments to produce a more representative dataset. Furthermore, aside from point clouds, the radar sensor is also capable of producing micro-doppler spectrograms. Other studies [2,10,32,34] have successfully utilized this data type thus, future work can explore the combination of both data types. The preprocessing algorithm and the deep learning model also have parameters that can be varied and explored, in addition to trying entirely different model architectures such as 3D CNNs or Transformers. Finally, expanding the scope to translating sentences and conversations is most desirable in the near future.

References

1. Senate Bill No. 2117, An act requiring the use of Filipino Sign Language insets for local news programs, amending for the purpose Section 22 of Republic Act No. 7277, as amended, otherwise known as the Magna Carta for Persons with Disabilities (PWDs) (2014). https://legacy.senate.gov.ph/lisdata/1868815815!.pdf

2. Adeoluwa, O., Kearney, S., Kurtoglu, E., Connors, C., Gurbuz, S.: Near real-time ASL recognition using a millimeter wave radar, p. 43 (2021). https://doi.org/10.1117/12.2588616
3. Al-Hourani, A., et al.: Chapter 7 - millimeter-wave integrated radar systems and techniques. In: Chellappa, R., Theodoridis, S. (eds.) Academic Press Library in Signal Processing, vol. 7, pp. 317–363. Academic Press (2018). https://doi.org/10.1016/B978-0-12-811887-0.00007-9, https://www.sciencedirect.com/science/article/pii/B9780128118870000079
4. Balbin, J.R., et al.: Sign language word translator using neural networks for the aurally impaired as a tool for communication. In: 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), pp. 425–429. IEEE (2016)
5. van Berlo, B., Elkelany, A., Ozcelebi, T., Meratnia, N.: Millimeter wave sensing: a review of application pipelines and building blocks. IEEE Sens. J. **21**, 10332–10368 (2021)
6. Cabalfin, E.P., Martinez, L.B., Guevara, R.C.L., Naval, P.C.: Filipino sign language recognition using manifold projection learning. In: TENCON 2012 IEEE Region 10 Conference, pp. 1–5. IEEE (2012)
7. Dong, Y., Yao, Y.D.: Secure mmWave-radar-based speaker verification for IoT smart home. IEEE Internet Things J. **8**(5), 3500–3511 (2021). https://doi.org/10.1109/JIOT.2020.3023101
8. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of 2nd International Conference on Knowledge Discovery, pp. 226–231 (1996)
9. Gurbuz, S., et al.: American sign language recognition using RF sensing. IEEE Sens. J. **21**, 3763–3775 (2020). https://doi.org/10.1109/JSEN.2020.3022376
10. Gurbuz, S.Z., et al.: American sign language recognition using RF sensing. IEEE Sens. J. **21**(3), 3763–3775 (2021). https://doi.org/10.1109/JSEN.2020.3022376
11. Hurlbut, H.M.: Philippine signed languages survey: a rapid appraisal (2008)
12. Iovescu, C., Rao, S.: The fundamentals of millimeter wave radar sensors. https://www.ti.com/lit/wp/spyy005a/spyy005a.pdf
13. Liddell, S.K., Johnson, R.E.: American sign language: the phonological base. Sign Lang. Stud. **64**(1), 195–277 (1989)
14. Liu, H., et al.: Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 4, no. 4, pp. 1–28 (2020)
15. Liu, H., et al.: M-gesture: person-independent real-time in-air gesture recognition using commodity millimeter wave radar. IEEE Internet Things J. **9**, 3397–3415 (2021). https://doi.org/10.1109/JIOT.2021.3098338
16. Liu, H.: Chapter 5 - autonomous rail rapid transit (art) systems. In: Liu, H. (ed.) Robot Systems for Rail Transit Applications, pp. 189–234. Elsevier (2020). https://doi.org/10.1016/B978-0-12-822968-2.00005-X
17. Magistrado, J.: Gloves na kayang mag-convert ng ph sign language sa boses, binuo ng ilang estudyante (2021). https://news.abs-cbn.com/news/07/02/21/sign-language-gloves-students-camsur
18. Martinez, L.B.: Observations on regional variants and handshape patterns of six signs in Filipino sign language (2009)
19. Meng, Z., et al.: Gait recognition for co-existing multiple people using millimeter wave sensing, vol. 34, pp. 849–856 (2020). https://ojs.aaai.org/index.php/AAAI/article/view/5430

20. Montefalcon, M.D., Padilla, J.R., Llabanes Rodriguez, R.: Filipino sign language recognition using deep learning. In: 2021 5th International Conference on E-Society, E-Education and E-Technology, pp. 219–225 (2021)
21. Movido, A.: Feeling left out, deaf community seeks government help to adjust to new normal (2020). https://news.abs-cbn.com/life/05/20/20/feeling-left-out-deaf-community-seeks-government-help-to-adjust-to-new-normal
22. Notarte-Balanquit, L.A.: Insights from the first Filipino Sign Language (FSL) summit & the prospects for Filipino sign linguistics (2021)
23. Oliva, K.E., Ortaliz, L.L., Tobias, M.A., Vea, L.: Filipino Sign Language recognition for beginners using kinect. In: 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), pp. 1–6. IEEE (2018)
24. Ong, C., Lim, I., Lu, J., Ng, C., Ong, T.: Sign-language recognition through gesture & movement analysis (SIGMA). In: Billingsley, J., Brett, P. (eds.) Mechatronics and Machine Vision in Practice 3, pp. 235–245. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76947-9_17
25. Palipana, S., Salami, D., Leiva, L.A., Sigg, S.: Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **5**(1), 127 (2021). https://doi.org/10.1145/3448110
26. Philippine Deaf Resource Center, I., Philippine Federation of the Deaf, I.: An Introduction to Filipino Sign Language, vol. 1: Understanding Structure. Philippine Deaf Resource Center, Inc. (2004)
27. Ren, Y., Lu, J., Beletchi, A., Huang, Y., Karmanov, I., Fontijne, D., Patel, C., Xu, H.: Hand gesture recognition using 802.11ad mmWave sensor in the mobile device. In: 2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), pp. 1–6 (2021). https://doi.org/10.1109/WCNCW49093.2021.9419978
28. Rivera, J.P., Ong, C.: Recognizing non-manual signals in Filipino Sign Language. In: Proceedings of Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 1–8 (2018)
29. Salami, D., Hasibi, R., Palipana, S., Popovski, P., Michoel, T., Sigg, S.: Tesla-rapture: a lightweight gesture recognition system from mmWave radar point clouds (2021)
30. Sandjaja, I.N., Marcos, N.: Sign language number recognition. In: 2009 Fifth International Joint Conference on INC, IMS and IDC, pp. 1503–1508. IEEE (2009)
31. Santhalingam, P.S., et al.: Expressive ASL recognition using millimeter-wave wireless signals. In: 2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), pp. 1–9 (2020). https://doi.org/10.1109/SECON48991.2020.9158441
32. Santhalingam, P.S., Hosain, A.A., Zhang, D., Pathak, P., Rangwala, H., Kushalnagar, R.: mmASL: environment-independent ASL gesture recognition using 60 GHz millimeter-wave signals. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 4, no. 1, pp. 1–30 (2020)
33. Wang, J., Ran, Z., Gao, Q., Ma, X., Pan, M., Xue, K.: Multi-person device-free gesture recognition using mmwave signals. China Commun. **18**(2), 186–199 (2021). https://doi.org/10.23919/JCC.2021.02.012
34. Wang, Z., Yu, Z., Lou, X., Guo, B., Chen, L.: Gesture-radar: a dual doppler radar based system for robust recognition and quantitative profiling of human gestures. IEEE Trans. Hum. Mach. Syst. **51**(1), 32–43 (2020)

Dehydration Scan: An Artificial Intelligence Assisted Smartphone-Based System for Early Detection of Dehydration

Priyeta Saha(✉), Syed Muhammad Ibne Zulfiker, Tanzima Hashem, and Khandker Aftarul Islam

Bangladesh University of Engineering and Technology, Dhaka 1205, Bangladesh
{1605094,1605110,1605063}@ugrad.cse.buet.ac.bd,
tanzimahashem@cse.buet.ac.bd

Abstract. Dehydration occurs due to fluid loss from the human body, affects regular body functions, and causes health complications. Physical exercises, poor fluid intake, and diseases like fever and diarrhea may result in dehydration. Current clinical and laboratory-based dehydration detection techniques are expensive, time-consuming, and require people to visit medical facilities, which often do not exist in destitute areas. Though recent research has focused on monitoring physiological parameters (e.g., heart rate, stress, and oxygen) and detecting diseases using smartphones, the area of dehydration detection has not been sufficiently addressed. We present a smartphone-based early dehydration detection system using artificial intelligence, which is ubiquitous, quick, and does not require any additional cost or expertise to operate. We develop a siamese network-based deep learning model to detect the changes in the facial landmarks that appear from dehydration and are not detectable with the naked eyes of general people. Our model provides an overall accuracy of 76.1% and is lightweight enough to run on a smartphone processor. By integrating it in the background, we develop a smartphone app, "Dehydration Scan" that simply captures facial images of individuals and detects their hydration status. Knowing early about dehydration allows people to take oral rehydration solutions and avoid severe dehydration.

Keywords: Dehydration detection · Mobile image analysis · Deep learning

1 Introduction

Dehydration occurs when the level of fluid in the human body falls below a certain threshold. Dehydration is a common effect of many diseases like fever,

---

P. Saha and S. M. I. Zulfiker—Both authors contributed equally to this research.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 289–312, 2023.
https://doi.org/10.1007/978-3-031-34586-9_20

diarrhoea and cholera, making it a significant cause of death worldwide. Lack of clean drinking water and proper sanitation causes people, especially in developing countries, to get infected with waterborne diseases like diarrhoea and cholera. Most people do not realize that they are dehydrated until it is too late. By then, hospitalization becomes a requirement. However, most of these people do not have economic solvency, nor do the hospitals have adequate human resources to provide proper healthcare in impoverished areas. Thus an increase in the death count becomes inevitable. Although these circumstances are more prevalent in lower-income countries, they exist in the developed world as well. People can also undergo mild to moderate levels of dehydration due to poor fluid intake or physical exercise. Studies [2,12] show that mild dehydration has an impact on alertness, concentration, mood, and cognitive abilities. Considering all the direct and indirect effects of dehydration on the human body and mind, a readily available diagnostic solution for the masses is a necessity at this point.

Compared to other mobile healthcare domains [11,21,22,32], researchers have not explored the topic of dehydration detection as much. Current dehydration detection techniques are primarily based on laboratory tests (blood or urine sample analysis) and clinical assessments undertaken by health professionals. Laboratory tests are expensive, time-consuming, and require specialized equipment. World Health Organization's IMCI algorithm [24] and DHAKA score [17] provide guidelines to health workers to detect dehydration based on clinical signs (e.g., sunken eye, skin turgor) and symptoms (e.g., thirst, pulse). However, health workers often do not have sufficient skills to follow the guidelines accordingly [1]. To reduce the need for skilled health workers or expensive medical equipment, researchers have developed dehydration detection techniques through kinetic analysis of hemoglobin concentration [33], exploiting photoplethysmographic signals [25] and measuring skin conductance [18]. However, they all require specialized hardware, making them expensive and inaccessible to the masses.

Dehydration causes very subtle changes to our facial landmarks, especially in regions like eyes and lips (e.g., reduced skin turgor or elasticity, tired and dry eyes, dry lips) [3,7,27]. These changes are not often perceivable by human eyes, especially when a person is in the early stages of dehydration. Therefore, it often goes unnoticed despite being a prevalent condition. We utilize artificial intelligence to work around the limitations of human vision in this case. On the other hand, with the widespread use of smartphones containing high-resolution cameras, mobile image analysis has emerged as a convenient solution for medical diagnosis [19,21,22]. We develop a smartphone-based early dehydration detection system using artificial intelligence that overcomes the limitations of the existing solutions:

- *Accessible* - Medical facilities or specialized hardware are not equally available worldwide, whereas smartphones are accessible almost everywhere.
- *Quick* - Our smartphone-based application provides instantaneous results, whereas diagnostic tests at hospitals usually take one or more days to deliver reports.

- *Automated* - Our smartphone-based detection tool uses artificial intelligence to detect dehydration and does not depend on the skill of the health workers for observing clinical signs.
- *Easy to Use* - The built-in high-resolution cameras in modern-day smartphones make the detection straightforward for users.
- *No Cost* - Using our smartphone-based detection system, patients can make initial or even intermediate-level assessments without taking expensive and invasive diagnostic tests at hospitals or buying specialized hardware.
- *Early Detection* - Early detection using smartphones can prevent further deterioration of a patient's condition. Mild dehydration detection using our application can save users from reaching a state where they need hospitalization.

We develop a deep learning model to detect the changes in facial landmarks when an individual becomes dehydrated. A traditional classifier [6,14,26] based on facial images would not perform well since the changes caused by mild dehydration are minute and can vary from person to person. For example, the dehydrated face of an individual with dry skin is not the same as that of an individual with oily skin. Similarly, the dehydrated face of an older adult is different from that of a young person. To overcome this challenge, we adopt the siamese neural network [31], proposed for a context similar to ours. Our model takes two images as input: one hydrated facial image as the reference and the other facial image representing the current state (hydrated or dehydrated) of the individual. The model predicts the class for the current facial image as hydrated or dehydrated. In Sect. 5, we elaborate on how we adopt the siamese network-based contrast learner with an appropriate loss function to produce outcomes for different facial landmarks and full facial image and then ensemble them to derive the final result (hydrated/dehydrated). Section 6 presents the performance of our model. Our model outperforms the baseline solutions by a large margin. Since our model, trained and tested on a dataset for mild dehydrated conditions, can detect dehydration with reasonably good accuracy, it is expected that our model will lead to even more promising results for moderate or severe dehydration cases.

One of the major challenges we face is the lack of a publicly available dehydration dataset that fits our needs. To mitigate this issue, we build our own dataset consisting of 2340 sample pairs of images (hydrated-hydrated or hydrated-dehydrated) from 70 healthy volunteers. We do not include volunteers who have any other disease in the dataset to eliminate the inference of other diseases on the facial landmarks. Research [16,18] shows that people usually have mild dehydration during fasting. Therefore we consider the month of Ramadan for data collection when practising Muslims fast from sunrise to sunset. We develop a data collection app to capture the facial images of the volunteers in hydrated and dehydrated states. Then we apply our preprocessing techniques to remove the image noises, improve the image quality and extract the images of the facial landmarks. The landmarks that show the most prominent changes upon dehydration are chosen, including the lips, eyes, and surrounding regions. The preprocessed images of different landmarks constitute our final dataset on which we train and test our model. Our data collection and preprocessing techniques are discussed in Sect. 3 and Sect. 4, respectively.

Finally, we develop a smartphone-based application, "Dehydration Scan" that captures the facial image of an individual using the integrated camera and classifies the condition of the individual as hydrated or dehydrated by running our developed model in the background. To the best of our knowledge, this is the first non-invasive approach for dehydration detection using a smartphone without requiring any additional equipment or expert skill while achieving adequate accuracy. In Sect. 7, we present our smartphone application to detect dehydration.

In summary, the contributions of this paper are as follows:

- We build a dataset that includes 2340 sample pairs of images from 70 individuals.
- We develop preprocessing steps to extract specific facial landmarks from the image frames and remove the noise associated with the effect of different lighting and background settings.
- We propose our siamese network-based contrast learning model to find the differences between a user's hydrated and dehydrated landmark images. We incorporate the individually derived values for different landmarks and the full facial image into a final score and classify the user's state as hydrated or dehydrated accordingly. We show the effectiveness of our model in experiments.
- We design and develop a complete and functional mobile application where users can take pictures of their faces and get to know their hydration status instantaneously.

2 Related Work

To date, existing works done on dehydration detection have been based on information collected and processed manually. In fact, no significant research has been done on automating this process. The existing methods require significant human intervention and often rely on a certain level of medical expertise. None of them focus on detecting dehydration using only mobile camera images. Table 1 shows a comparative analysis of existing dehydration detection techniques with ours in terms of different features. We observe that only our solution supports all desirable features.

The World Health Organization developed the Integrated Management of Childhood Illness (IMCI) algorithm [24] to guide health workers in detecting different levels of dehydration by monitoring clinical symptoms. Levine et al. [17] came up with the Dehydration: Assessing Kids Accurately (DHAKA) score for dehydration detection based on clinical symptoms. The authors conducted a study in Bangladesh, where local nurses analyzed children's dehydration status using both the DHAKA score and the IMCI algorithm. In a detailed comparison of the results, the DHAKA score was the more accurate predictor of dehydration for children. In [4], the authors developed a smartphone application that takes

Table 1. A Comparative Analysis of Existing Dehydration Detection Solutions

Paper	Technique	Health professional	Additional equipment	Additional cost	Time consuming
[4, 17, 24]	Clinical symptom monitoring	✓	✗	✓	✓
[18]	Skin conductance monitoring	✗	✓	✓	✗
[23]	Remote optical monitoring	✓	✓	✗	✗
[20]	Sweat electrolyte conductance monitoring	✗	✓	✓	✗
[25]	Oximeter signal monitoring	✗	✓	✓	✗
[19]	Image processing to detect skin mechanical properties through skin turgor test	✓	✗	✓	✓
Ours	Image processing using facial landmarks	✗	✗	✗	✗

clinical symptoms as input and then applies the IMCI algorithm of WHO to produce the result. This work aimed to check whether shifting from paper-based work to the smartphone app can improve the reliability and usability of the dehydration detection system.

The diagnostic technique proposed in [18] uses a non-invasive wearable sensor for collecting skin conductance data and detects dehydration based on the skin conductance state. Another study [10] examined the correlation between dehydration severity and the moisture level of oral mucous membrane measured through a moisture-checking device. In [33], the authors hypothesized whether dehydration can be detected using kinetic analysis of hemoglobin concentration.

In [23], the authors used a wristwatch to extract several bio-medical parameters. They implemented two optical approaches. One of them was the rotation of linearly polarized light by certain materials exposed to magnetic fields. Another was the extraction and separation of remote vibration sources. In [20], the authors used a conductometric sensor to measure sweat electrolyte conductance to detect dehydration. Another study [25] used photoplethysmographic signals with small, wearable pulse oximeters and set features based on the variable frequency complex demodulation. Those features were then fed to a support vector machine model to detect dehydration.

In [19], the authors used smartphones to capture the videos of skin turgor tests done on hands with two different methods - skin mark method and skin texture method. They ran image processing algorithms to extract skin mechanical properties and hydration levels from the frames of those videos. They used smartphones to track the turgor test's skin stretching and relaxation processes, which medical professionals usually undertake to check for dehydration.

3 Data Collection

At present, there exists no image dataset of people suffering from dehydration. Since our target is to detect dehydration in the early stages and people who come to the hospital are severely dehydrated in most cases, collecting data from hospitalized patients would not serve our purpose. To work around this problem, we leveraged the month of Ramadan, when practising Muslims fast from sunrise to sunset. Research [16,18] shows that people usually have mild dehydration during fasting, which does not result in severe health issues.

We developed a mobile application to achieve two goals:

1. To streamline the data collection process, and
2. To enable end-users to evaluate their hydration status instantaneously.

3.1 Data Collection Mobile Application

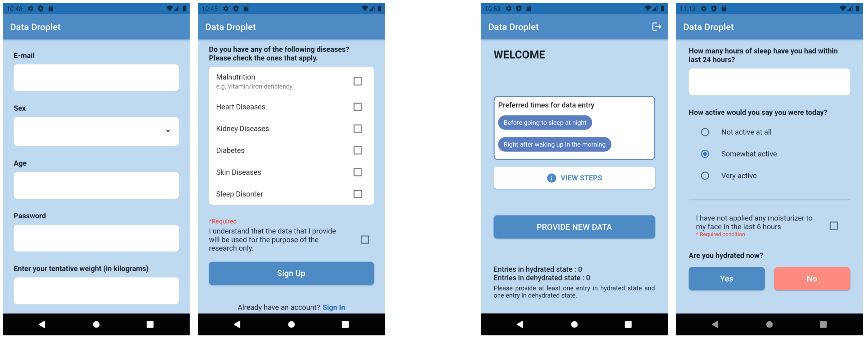
Our aim was to build an application that works on any mobile device irrespective of the operating system and configuration. Thus, we chose Flutter [30], a cross-platform application development framework, to create the smartphone app. We also used the Firestore Database and Storage services of Firebase [29] as the system's backend.

The app included five well-defined steps for every user to follow. Figure 1 shows a detailed view of the application interface in these steps.

1. Users created a new account by providing relevant information, e.g., email address, age, sex, weight, and existing health conditions on the first usage. They were asked for explicit permission regarding the use of their data for our research before account creation. Afterward, they could log into their accounts from any smartphone device.
2. Every time users opted to provide data through the app, they had to give some basic information first, such as hours of sleep and activity level. Using moisturizer on the skin can neutralize the signs of dehydration almost entirely. So the app also required them to confirm that they had not applied any moisturizing product on their face in the last 6 h prior to providing the entry. We instructed the users to select hydrated state if they provide data at night after fluid intake and to select dehydrated state if they provide data during fasting.
3. Next, the users had to choose hydrated or dehydrated as their current state, which we later validated based on additional information. If hydrated, they were asked to mention their amount of fluid intake (in glasses) in the last six hours. Otherwise, they were asked to pick the approximate time of their last fluid intake. Based on the answer to the additional question, we verified the correctness of the user state (hydrated or dehydrated).
4. In the most crucial step, users recorded 5-second long videos of their faces using the front or back camera. The app displayed a bounding box around the users' faces in this process to ensure that no part of the face was outside

the camera view. They were also given the option to preview their captured face video and retake it if necessary before uploading it to Firebase cloud storage.

5. Upon successfully uploading a face video to Firebase cloud storage, the app stored the upload time against the user's credentials in its local storage. We implemented this additional check to disable the option for that user to give two data entries in less than 6 h gap.



E-mail

Sex

Age

Password

Enter your tentative weight (in kilograms)

Do you have any of the following diseases? Please check the ones that apply.

Malnutrition e.g. vitamin/iron deficiency ☐
Heart Diseases ☐
Kidney Diseases ☐
Diabetes ☐
Skin Diseases ☐
Sleep Disorder ☐

*Required
I understand that the data that I provide will be used for the purpose of the research only. ☐

Already have an account? Sign In

Preferred times for data entry
Before going to sleep at night
Right after waking up in the morning

Entries in hydrated state : 0
Entries in dehydrated state : 0
Please provide at least one entry in hydrated state and one entry in dehydrated state.

How many hours of sleep have you had within last 24 hours?

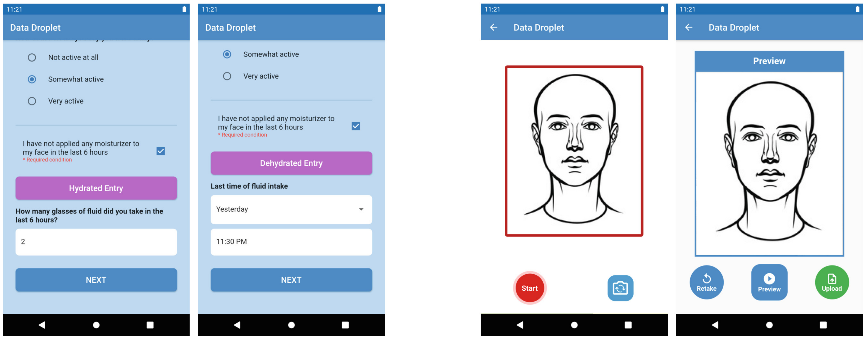
How active would you say you were today?
○ Not active at all
● Somewhat active
○ Very active

I have not applied any moisturizer to my face in the last 6 hours ☐
* Required condition

Are you hydrated now?

(a) Account creation with basic information and consent

(b) Basic information and regulation during data entry



○ Not active at all
● Somewhat active
○ Very active

I have not applied any moisturizer to my face in the last 6 hours ☑
* Required condition

How many glasses of fluid did you take in the last 6 hours?
2

● Somewhat active
○ Very active

I have not applied any moisturizer to my face in the last 6 hours ☑
* Required condition

Last time of fluid intake
Yesterday
11:30 PM

(c) Additional information based on hydrated/dehydrated state (for dehydrated entry, the user can choose between today and yesterday and select the time from a time picker dialog)

(d) 5-second face video capture with preview and retake options (a dummy image of a human face has been shown to protect the privacy of the users who provided data)

Fig. 1. Various steps of the data collection portal of our smartphone application

The app supports a minimum SDK version of 21, which is the earliest release of the Android SDK that it can run on. It is equivalent to Android 5.0 (API level 21) or higher. The fundamental features and functionalities of Android are available in this version and all subsequent versions. Since the latest stable version is Android 10.0 (API level 29), our application is runnable on a wide range of configurations. We used a few basic modules in our application - camera (for recording face videos), video player (for displaying a preview of captured video), and shared preferences (for locally storing entry timestamps) which work smoothly in most smartphones.

Table 2. Age distribution of dataset subjects

Age Range	No. of Participants
10–20	8
21–30	45
31–40	13
41–50	4

3.2 Dataset

We developed and distributed the mobile application described previously to build an in-house face video dataset explicitly suited to our use case. We prepared a final dataset containing 2340 pairs of images by strategically extracting and combining frames from the collected videos. We had a total of 70 volunteers in this process. 36 of them were male, and 34 were female. The participants were between 10 and 50 years old. However, many of them were undergraduate students, which added an age bias to our dataset. The age distribution is given in Table 2.

Initially, we had 326 face video entries from 70 individuals, which means that, on average, every user contributed 4–5 entries to our dataset. We collected videos instead of images because we could extract up to 10 frames from each video entry. To put it simply, we extracted a maximum of 10 hydrated or 10 dehydrated images of the same user from each video of that user. We discarded blurry or unusable frames using some preprocessing steps. Nevertheless, this approach helped us immensely to generate more data points for training and testing our model. Additionally, we checked if we had at least one hydrated and at least one dehydrated video entry from every user. Since we needed images of both conditions to create pair inputs for our model, we discarded the entries with no corresponding entry for the other condition. To ensure that our dataset does not reflect the symptoms of any disease or condition other than dehydration, we utilized the information gathered from the app. During account creation, the app displayed a list of diseases (e.g., malnutrition, heart diseases, kidney diseases, diabetes, skin diseases, sleep disorder) that can interfere with the signs of dehydration. From the list, users selected any condition that applied to them. While preparing our image dataset, we excluded all users suffering from any of the mentioned diseases. This filtering was necessary to prevent our model from falsely identifying symptoms of those diseases as effects of dehydration.

Ground Truth Validation. In the case of dehydrated data entries, we collected the number of hours passed after the last fluid intake as additional information from the user and used it to validate the given hydration status label. We found 11 video entries for which this time gap was less than 6 h. Thus, we discarded those noisy entries considering that the users might not be in a dehydrated state in those and proceeded with the remaining 315 entries. The average number of hours after the last fluid intake in the selected entries was 12 h. On the other hand, for hydrated entries, we collected the amount (in glasses) of fluid consumed in the last 6 h as additional information to make sure that the user is in a hydrated state. The average glass count was 3.

For every user, we selected one of the hydrated images as the reference image for that user. Then we created all possible combinations by pairing the reference image with every available image (hydrated/dehydrated) of that user. While generating these combinations, we did not consider the creation time or sequence of a particular entry. The only condition in pairing any hydrated or dehydrated image with the reference image was that the two images belonged to the same user. We conducted this step of the process based on the assumption that for any particular user, the effects of dehydration on different facial regions would be pretty similar regardless of the time he/she captured the condition. In this step, we significantly augmented the face video dataset collected through our app and prepared a final dataset of 2340 pairs of images. The dataset is balanced, consisting of 1170 reference-hydrated and 1170 reference-dehydrated pairs of samples.

4 Data Preprocessing

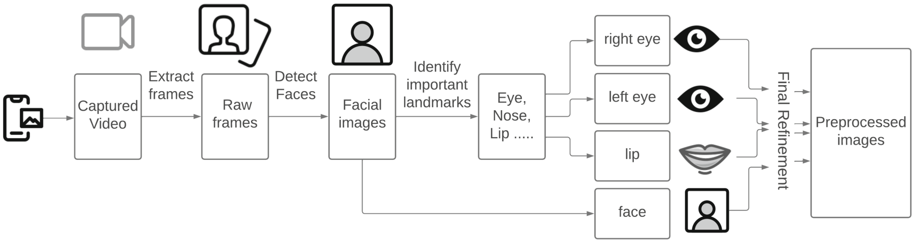


Fig. 2. Preprocessing steps

In this section, we elaborate on the automated preprocessing steps performed before we feed our data into the model for training and prediction. Our model is at its core a siamese network that uses contrastive loss in order to learn and detect the differences between a hydrated and a dehydrated image. Thus, we need to provide pairs of images of the same individual to the model during training. The model calculates a similarity score for the new image to predict its hydration level using the trained weights.

From the data collection portal of our app, we obtain 10-second long videos of both hydrated and dehydrated states of each user. These videos have a lot of noise, such as varying levels of light in the image, differences in resolution, contrast, saturation, and overall image quality due to variation in image sensors. We need to eliminate these adverse effects and format the final image to emphasize the maximum contrast between the hydrated and dehydrated states and offset the other differences. Since our model trains with images, the first step is to extract individual frames from the videos. We do this using the python library OpenCV. Using the variance of the Laplacian, we filter out the blurry images and keep the ones that report the highest values of sharpness. We select only the images that cross a certain variance threshold and discard the rest. A final manual pass is done to ensure that no anomalous images have spilled through. An overview of the preprocessing steps is presented in Fig. 2.

Our final model uses the similarity scores of facial landmarks as different features and creates an ensemble by assigning different weights to each of these scores. We determine the combination of weights for which our model provides the best performance in experiments. The values of the weights that our model uses for different landmarks are specified in Sect. 6.1. We train individual models for each of the landmarks separately. Thus, it is paramount for us to identify, detect and segregate the landmarks in our preprocessing step. Since signs of dehydration are primarily visible in areas surrounding the eyes, the skin, and lips, we particularly emphasize our landmarks in those areas. For the eyes, we include areas around the eyelids as well as the under-eye region since those are the areas most affected. The entire face image is passed on as a separate landmark as well.

We use the Haar Cascade classifier included in OpenCV to extract the landmarks from images. The classifier returns four coordinates for each of our individual landmarks. Each of the four coordinates represents a corner of a bounding box surrounding the landmark. Joining the four coordinates, we obtain a bounding box locating the position of the landmark. We then separate the landmarks by cropping them out from the original image. Lastly, we scale the images to our required pixel size of 200*200 before feeding them into the model. The pixel size is chosen to strike the optimal balance between retaining the highest possible level of details while keeping it lightweight enough for on-device inference.

5 Model Overview

Unlike most facial image classification problems, which are typically solved by utilizing model-extracted features, dehydration detection is incredibly challenging because the symptoms of mild dehydration are tough to decipher. Therefore, simply attempting to capture facial features with a traditional deep learning model would not perform well in this particular problem.

We extract and prepare images of specific facial landmarks in the preprocessing step, which is quite crucial to make it easier for our model to capture differences down to the tiniest detail. Our system requires only one reference image

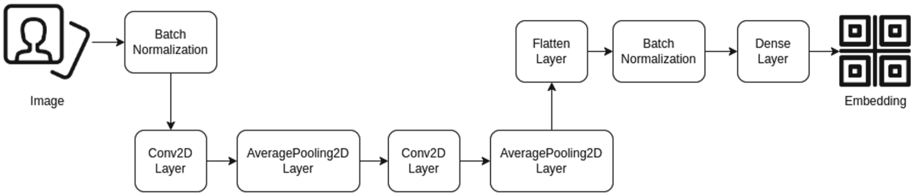


Fig. 3. Layered view of base convolutional network

taken in the hydrated state from every individual. Thus, we build a model that can predict a user's hydration status from a pair of images - one of them is the initial reference image captured in the hydrated state, and the other is an image taken at any hydration level at any moment. To serve this purpose, we leverage an artificial neural network model called siamese network [5], specifically designed for use cases similar to ours. Koch et al. [15] used siamese neural networks to address the one-shot classification problem where predictions need to be made from a single instance of any class. The way we structure our approach is quite similar to this problem. Given a reference hydrated image of any user, our model processes any input image of the same user, evaluates its similarity to the reference, and classifies it as hydrated or dehydrated. Since the model only requires a single pair of images from any user, its prediction technique is comparable to one-shot learning.

The siamese network-based architecture primarily comprises two identical towers or sister networks with shared weights. Each takes one image as input and generates a feature embedding for that image.

In Fig. 3, we can see a detailed graphical overview of our convolutional network. Each network is essentially a two-dimensional convolutional neural network containing the same sequence of layers. The input images first go through batch normalization before getting processed by the convolutional networks. Our model constructs the networks from consecutive 2D convolutional and average pooling layers, followed by batch normalization and a dense layer with rectified linear unit (ReLU) activation. Our model then converts the generated feature embeddings into a single merged layer by computing their cosine distance.

$$similarity(A, B) = \frac{A \cdot B}{\|A\|_2 \times \|B\|_2} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

As can be seen from Eq. 1, the cosine similarity of two vectors A and B is simply the normalized dot product of A and B. Here, $\|A\|_2$ is the L2 norm of vector A. A high cosine similarity implies that the images closely match each other and have minimal contrast between them.

$$distance(A, B) = 1 - similarity(A, B) = 1 - \frac{A \cdot B}{\|A\|_2 \times \|B\|_2} \tag{2}$$

From Eq. 2, we can see that cosine distance and cosine similarity are complementary measures. Two identical vectors with an angle of zero degrees between them have a similarity score of 1 and a distance of 0.

Using other distance metrics such as Euclidean distance and Manhattan distance in the merging step does not give promising results. The Euclidean distance between two vectors refers to their straight-line distance, while the Manhattan distance refers to the sum of the distances along each axis. Unlike these measures, cosine distance depends on the angle between two vectors and does not consider the size of the vectors. Thus, in our case, cosine distance proves to be a better metric for differentiating the embeddings. The distance values that signify the contrast between the input images are passed to the final neural network consisting of batch normalization and a typical dense layer with sigmoid activation. To summarize, our proposed contrast learning model takes pairs of images belonging to the same user as input, estimates their similarity, and produces a score between 0 and 1 accordingly. The architecture of this model is illustrated in Fig. 4.

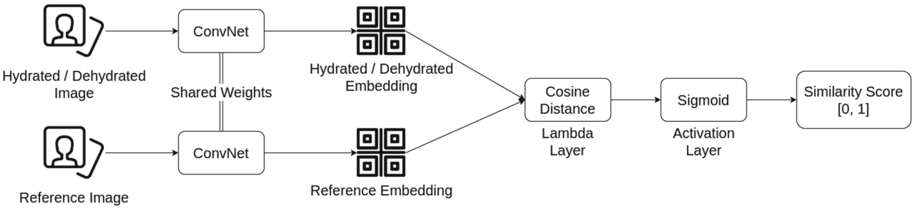


Fig. 4. Siamese-network based contrast learner

Another critical aspect of the model is the choice of the loss function, which is used in the backpropagation to update the weights of the convolutional layers. Two loss functions are generally used in siamese networks - triplet loss and contrastive loss [13]. If triplet loss is used, the model takes in three inputs - one reference image, one similar or neighbor image, and one dissimilar or distant image. The key idea in this approach is to minimize the reference-neighbor distance and maximize the reference-distant distance. Contrastive loss, on the other hand, deals with pairs of images. The pairs can be reference-neighbor or reference-distant, which is the exact format of the samples in our dataset. Thus, we choose contrastive loss as the loss function for our model.

$$L = Y \times D^2 + (1 - Y) \times \max(margin - D, 0)^2 \tag{3}$$

In Eq. 3, the contrastive loss L is computed from the predicted value Y, the cosine distance D, and the baseline distance *margin* for which the model should classify pairs as dissimilar. The default value for the *margin* is 1.

The contrast learner is run parallelly for the image pairs of each of our selected facial landmarks - the right and left eyes, lip, and nose. We acquire a similarity score from every execution of the model and calculate a weighted sum of those scores to derive the final prediction. The weights for the landmarks are obtained through an extensive trial and error process. After adequate experimental analysis, we find nearly optimal weights to account for the relevance of each landmark in maximizing the prediction accuracy. We also run the model on entire face images extracted from video frames. In short, to reach the conclusive prediction of our system, we take into account every essential region of the face, do individual similarity estimations using our siamese network-based model, and combine them to classify the input image as hydrated or dehydrated. Figure 5 shows an overview of our proposed model.

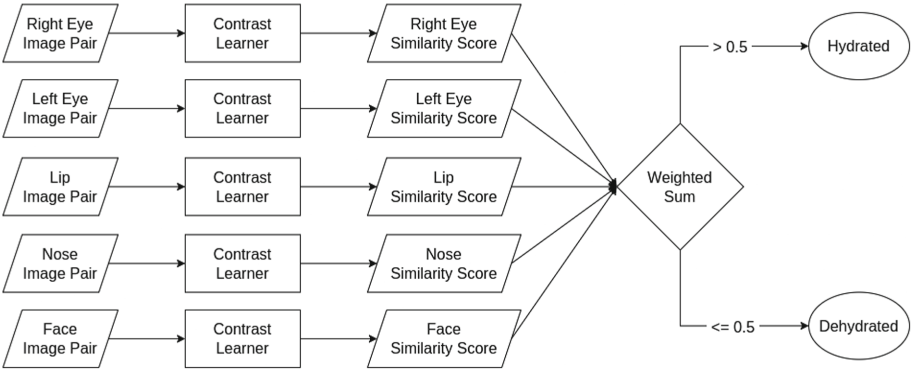


Fig. 5. Classification using a weighted sum of similarity scores from image pairs of individual landmarks

There are two core factors behind the considerable prediction accuracy of our proposed model.

- We take a personalized approach instead of generalizing the problem. We focus on the contrast between hydrated and dehydrated images of the same person, which helps us cut out the differentiating factors among separate individuals.
- We run our model separately on the segregated and preprocessed images of various facial regions to better capture the changes occurring in each region. This approach is particularly substantial for identifying mild dehydration since the symptoms might not appear in the same region for every individual.

One additional concern in the development of our model is handling the resource constraints of smartphones. In order to predict hydration status using the limited processing power of smartphones, we choose TensorFlow as our artificial intelligence framework and devise a sufficiently lightweight model. We convert our TensorFlow model to a TensorFlow Lite model, which is smaller, significantly faster, and runs more efficiently on a mobile processor. Our smartphone application runs the lite model in the background whenever a user submits an input image of his/her face, performs on-device computation, and delivers instantaneous results - hydrated or dehydrated.

6 Performance Analysis

In this section, we present the performance of our proposed dehydration detection model.

6.1 Experimental Setup

Dataset and Preprocessing: We train and evaluate our model on our procured dataset elaborated in Sect. 3. The data first passes through some preprocessing steps as discussed in Sect. 4 before being split into train, test, and validation sets. As mentioned in Sect. 3.2, our dataset contains 2340 pairs of images, where 1170 pairs contain two hydrated images, and the other 1170 pairs include one hydrated and one dehydrated image. We take 60%, 20%, and 20% of the pairs as train, validation, and test sets, respectively. We use samples of three separate sets of users for training, validating, and testing datasets. The train, validation, and test split is discussed more elaborately in Sect. 6.6.

We calculate all performance metrics presented in this section on our test set, which is completely isolated from the training phase with no peeking involved anywhere in the pipeline.

Parameter Settings: We select a batch size of 16 and run the model for 10 epochs. The number of epochs is chosen through experimentation as elaborated in Sect. 6.5. The image size is selected to be 200*200 pixels in order to capture as much detail as possible while keeping it reasonably less resource-heavy. For the final ensemble, we choose weights for each of the individual landmarks as follows: Left and Right Eye: 0.275, Lip: 0.225, Nose: 0.125, Entire Face: 0.1.

Evaluation Criteria: We choose Accuracy, Specificity, Recall, and F1 score as our evaluation criteria. Accuracy measures how many hydrated and dehydrated images are correctly classified as hydrated and dehydrated, respectively. Specificity measures out of all the dehydrated images, how many are correctly classified as dehydrated. Recall tells us how many are correctly classified as hydrated from all the hydrated images. To understand the F1 score, we first need to know what precision is. Precision measures how many images that are classified as hydrated are actually hydrated. F1 score is the harmonic mean of precision and recall.

6.2 Comparative Analysis with Baselines

We consider the following four baselines and compare the performance of our solution with them.

- **Basic CNN:** For our basic CNN structure, we have used one convolutional layer followed by a pooling layer and a dense layer at the end. This is representative of the most rudimentary implementation of an image classifier.
- **VGG16** [26]**:** VGG was invented with the purpose of enhancing classification accuracy by increasing the depth of the CNNs. VGG 16 has 16 weight layers and is used for object recognition.

- **Xception** [6]**:** Xception is a convolutional neural network architecture consisting of 71 layers. This network gives a rich representation of images when trained on an expansive dataset.
- **Resnet50** [14]**:** Resnets are neural networks that use skip connections to solve the problem of vanishing gradients. Resnet50 is a Resnet that is 50 layers deep.

For VGG16, Xception, and Resnet50, we use transfer learning by initializing the models with weights trained on imagenet [8]. After that, all the baselines have been run on our dataset end-to-end to come up with the final predictions. The performance metrics are then compared with those of our proposed solution. We can get an overview of the performance comparison between the baselines and our model from Table 3.

Table 3. Performance comparison of our proposed model with different baselines

Model	Accuracy	Specificity	Recall	F1 score
Basic CNN	51.1%	15.2%	48.4%	56.6%
VGG16	34.4%	31.9%	52.3%	52.6%
Xception	42.5%	17.6%	38.7%	47.1%
ResNet50	65.9%	23.5%	99.7%	79.3%
Our Model	76.1%	52.1%	99.1%	80.7%

It is apparent from the above data that our proposed siamese network-based model provides a better classification accuracy, specificity, recall, and F1 score than those of the baselines. The key differentiating factor here is the emphasis of our model on detecting contrasts between the hydrated and dehydrated states. To assist the contrast detection, we concentrate on areas of the face that show the most significant changes during dehydration. Our model can learn differences in facial features upon dehydration better than the conventional models, which look at the face image as a whole and do not prioritize specific landmarks like our model.

6.3 Effect of Landmarks

To make the contrast detection between hydrated and dehydrated states more robust, we identify the landmarks that show changes upon dehydration. We now discuss how each of these individual landmarks performs on its own to detect dehydration. The performance comparison of how each of the landmarks performs separately is presented in Table 4.

The left and right eyes show the best performance out of all the landmarks. This is expected since one of the main symptoms of dehydration is sunken eyes. This is consistent with the studies made in [28] where it is mentioned that

Table 4. Performance comparison between individual landmarks

Landmark used	Accuracy	Specificity	Recall	F1 score
Entire Face	58.8%	20.4%	97.2%	70.2%
Nose	59.7%	19.5%	99.7%	71.3%
Lip	61.9%	29.3%	94.4%	71.2%
Right eye	76.1%	52.1%	96.3%	80.7%
Left eye	72.3%	69.3%	75.3%	73.1%
Left eye flipped	74.3%	50.6%	97.9%	79.2%

dehydration symptoms like sunken eyes, undereye darkness, or discoloration can show up faster because the skin in this region is thinner than other body parts. To properly capture this, we took into account the regions surrounding the eyes during our initial crop. The next best landmark is the lip, most likely due to the flakiness or dryness visible when there is no fluid intake for an extended period. [9] cites dry lips as one of the leading signs of dehydration. The entire face image gives the lowest accuracy, which further proves why conventional models do not perform well in this scope. The entire face data taken together do not provide enough information for classifying a dehydrated image from a hydrated one.

Additionally, we flipped left eye images horizontally to make them similar to right eye images. We included this step to make all the eye images more homogeneous and minimize the unexpected performance gap between left and right eye images.

6.4 Ablation Study

We perform an ablation study by removing one landmark at a time and performing the classification. We intend to see the effect of each of the landmarks on the final classification. The results of the ablation study are summarized in Table 5.

It is evident from the data that excluding any landmark reduces the performance of our model; for example, the obtained accuracy for excluding any

Table 5. Ablation Study

Landmark excluded	Accuracy	Specificity	Recall	F1 score
Entire Face	70.6%	45.7%	95.4%	76.4%
Nose	70.9%	41.9%	95.6%	77.5%
Lip	68.3%	36.7%	94.3%	75.9%
Right eye	64.4%	28.8%	92.4%	73.8%
Left eye	65.7%	44.7%	97.8%	78.3%
Left eye flipped	65.3%	30.6%	95.8%	74.2%

landmark is less than the overall accuracy 76.1%. Thus, each of the landmarks that we use as model features contributes to improving the prediction. The effect of removing the eyes from the classification has the highest impact. This is followed by the removal of the lips. This is consistent with our assumption that the eyes and lips show the most visible changes during dehydration.

6.5 Effect of Number of Epochs

In this experiment, we choose the optimal number of epochs to get the best performance of the model. Usually, up to a certain threshold, the more epochs the model runs for, the better the accuracy and performance measures are. After a point, the model starts overfitting to the training set, and its performance starts degrading. We need to find the optimal epoch number that provides the best results without introducing overfitting. Table 6 and Fig. 6 give us an overview of the performance metrics by varying the number of epochs.

Table 6. Performance comparison between varying Epoch numbers

No. of Epochs	Accuracy	Specificity	Recall	F1 score
5	68.1%	36.3%	95.7%	75.8%
8	71.6%	43.3%	97.8%	77.9%
10	76.1%	52.1%	99.1%	80.7%
12	66.7%	39.1%	94.4%	73.9%

The data shows that the model reaches maximum performance at 10 epochs. After that, the performance slowly starts degrading as it starts introducing overfitting. We, therefore, stopped the model after 10 epochs during our final classification.

6.6 Effect of Train-Test Ratio

In this section, we present our experiments to choose the optimal split for the test, validation, and train sets. On the one hand, increasing the train set size gives our model more data to train on, essentially boosting the performance. However, keeping little data on the validation set runs the risk of overfitting the train set. To find the optimal balance, we conduct multiple runs with different ratios of the split as presented in Table 7.

As we can infer from the data, the optimal performance is achieved with a 60-20-20 split of train, validation, and test set.

6.7 Male vs Female Participants

In this section, we compare the performance metrics of male and female participants. The findings are presented in Table 8. The results for male participants

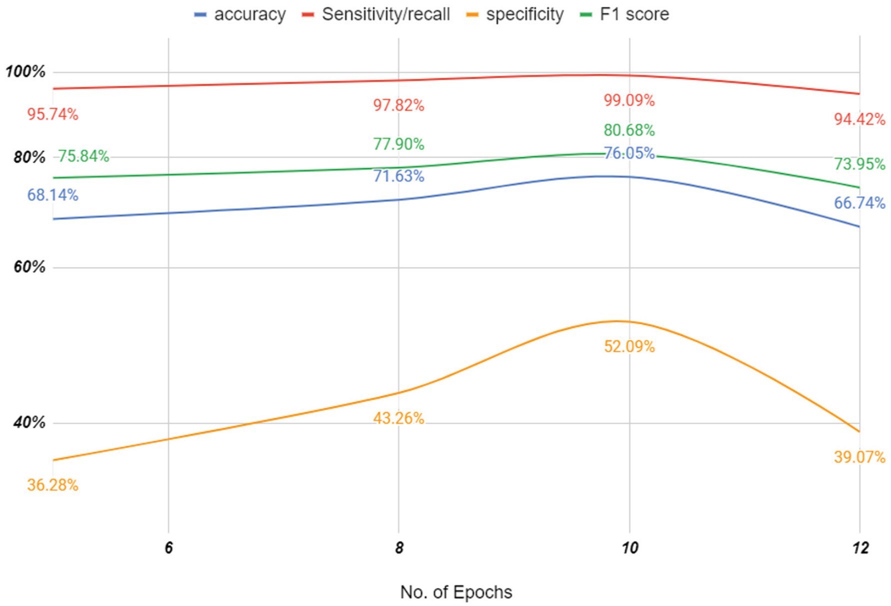


Fig. 6. Log graph of performance comparison between varying epochs

Table 7. Performance comparison between varying Train-Test ratios

% Train	% Validation	% Test	Accuracy	Specificity	Recall	F1 score
70%	15%	15%	62.1%	30.9%	93.4%	71.1%
70%	10%	20%	69.7%	48.1%	91.3%	75.1%
80%	10%	10%	54.3%	37.7%	71.1%	60.8%
60%	20%	20%	76.1%	52.1%	99.1%	80.7%

were derived by training and testing our model only on male data. We repeated the same process for female participants.

Although the accuracy is quite similar in both cases, there is a noticeable difference in the specificity and recall values. This indicates that our model performance might vary across genders.

6.8 Effect of Multiple vs Single Reference

Here we compare the performance of our model using single and multiple reference images. In the case of a single reference image, we paired the reference with each hydrated and each dehydrated image of the same individual. Similarly, in the case of multiple reference images, we selected multiple hydrated images for a user based on their sharpness and paired each of the references with each remaining hydrated and dehydrated image. The performance comparison is shown in Table 9.

Table 8. Performance comparison between male and female participants

Participant Gender	Accuracy	Specificity	Recall	F1 score
Male	72.9%	64.9%	99.8%	78.7%
Female	72.1%	83.6%	60.7%	68.5%

Table 9. Performance comparison between single and multiple reference image

No. of references	Accuracy	Specificity	Recall	F1 score
1	76.1%	52.1%	99.1%	80.7%
2	72.2%	44.4%	99.7%	78.3%
3	68.4%	36.8%	99.8%	75.9%

It is evident from the experiment results that using multiple images as references does not improve our model performance. One possible reason behind this might be that the hydrated frames of an individual are very similar to one another. Therefore, multiple reference images captured in hydrated conditions only increase the dataset size without contributing any new, substantial information regarding dehydration. Besides, the model might overfit the data of the users in the train set and, consequently, give a poor performance for face images of unseen test subjects.

6.9 Cross Validation

We partitioned the data into 5 non-overlapping folds or subsets and performed cross-validation by running our model 5 times. Every time, we chose a different subset for testing and used the remaining data for training and validation. The results from the different runs are shown in Table 10 along with the average performance.

Table 10. 5-fold cross-validation results

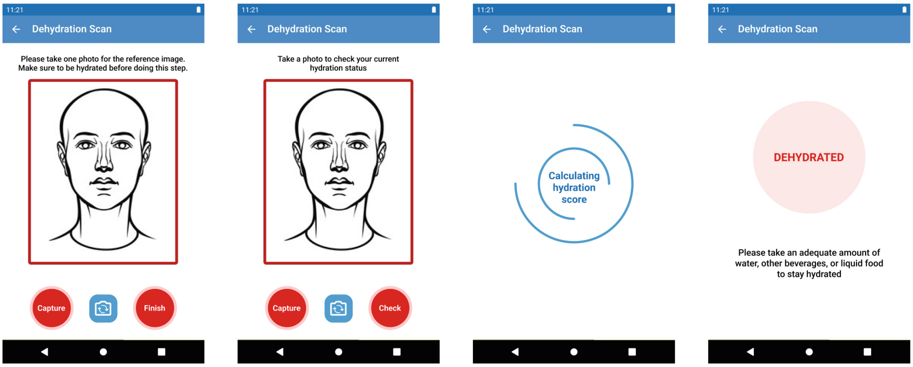
Test set	Accuracy	Specificity	Recall	F1 score
1	69.4%	59.7%	78.9%	72.1%
2	69.7%	45.6%	93.9%	75.6%
3	76.1%	64.9%	87.2%	78.5%
4	73.7%	47.4%	99.7%	79.2%
5	76.1%	52.1%	99.1%	80.7%
Average	73.0%	53.9%	91.8%	77.2%

7 “Dehydration Scan” App Overview

In this section, we present the interface and functionality of our smartphone application, “Dehydration Scan”. The app has been designed to minimize the number of steps that a user must complete to check his/her hydration status. The user interface is also relatively simple and highlights the required steps. We share a detailed view of the interface in Fig. 7.

While using the app, a typical user performs the following actions.

1. First, a user needs to take a picture in hydrated condition to start using the app. This picture is saved in the local storage of the app as the reference image for future predictions. Since the facial condition of the user may change over time because of factors like age and weather, the app provides an option to retake and update the reference image at any point in time. Besides, it is crucial for the user to be adequately hydrated while capturing the reference image. Otherwise, our dehydration detection model will fail to deliver accurate predictions.
2. Afterward, the user only needs to provide a face image to check for dehydration. This image goes through the preprocessing steps mentioned in Sect. 4 and gets paired with the previously captured and stored reference image to prepare the input for the model.
3. Then the app runs the trained lite version of our proposed model on the pair of images. This computation takes place entirely on-device and does not require any cloud storage or services. Since the model has already been trained on our dataset, it does not require any additional training time and executes in a matter of seconds.
4. After calculating the final score, the app displays the predicted class of the condition of the user.



Please take one photo for the reference image.
Make sure to be hydrated before doing this step.

Take a photo to check your current hydration status

Please take an adequate amount of water, other beverages, or liquid food to stay hydrated

(a) Taking a reference image (updatable) from the user for future evaluations

(b) Taking image input to pair it up with the reference image for comparison

(c) Running the on-device model to compute hydration score for the input image

(d) Displaying the final prediction to the user - either Hydrated or Dehydrated

Fig. 7. Steps of dehydration detection using Dehydration Scan

We use the TensorFlow Lite library to run our model in the background, and the minimum SDK version for this library is 21. Therefore, like our data collection application described in Sect. 3.1, Dehydration Scan also operates well in android devices with API level 21 or higher. This range of configurations covers almost all available android smartphones. Moreover, Dehydration Scan is a standalone smartphone application that does not require an internet connection to function. Thus, it serves to be a ubiquitous solution to the dehydration detection problem.

8 Limitations and Future Direction

While we demonstrate the strong potential of our proposed siamese model, there are scopes to improve it even further. Following is a list of limitations we identified from our research and the corresponding future research directions one might pursue:

- **Mitigating the absence of a publicly available diversified dataset:** There is an acute deficiency of a publicly available diversified dataset in the domain of dehydration detection. Although we have managed to procure a decent dataset of our own, it lacks diversity in ethnicity, age, and other aspects. Also, the dataset we worked on is imbalanced in terms of the age range of the participants. There are very few data points over the age of 40. Our dataset is primarily composed of undergraduate students and, as such, introduces an age bias. Overall there is massive scope for a better, more robust dataset. Improving the dataset can also improve the classification performance.
- **Automating the assignment of weights to the different landmarks:** In our implementation, we have manually assigned weights to the different landmarks during the computation of the final prediction. We have chosen the weights that provide the overall best performance over the entire test set through trial and error. However, the optimal weights may vary for each individual, so a generalized assignment of weight for everyone may not give the best results.
- **Using more sophisticated models with higher parameter counts:** As smartphone hardware gets more powerful year after year, it becomes feasible for more complex and sophisticated algorithms to be run on-device on mobile processors. The limitations in hardware capabilities will not be present in a few years when more powerful smartphones penetrate the general mass. Therefore, future research should concentrate on more sophisticated models that give higher prediction accuracy, albeit with more resources.

9 Conclusion

We developed a non-invasive smartphone-based dehydration diagnostic solution that does not need additional cost, hardware, or skilled workers and produces results instantaneously. Since our approach delivers acceptable accuracy (on average 76.1%) for mild to moderate dehydration, we can infer that more severe cases of dehydration can be predicted with a higher degree of confidence. Our developed siamese network-based dehydration detection model outperforms the baseline models by a large margin (i.e., our model achieves at least 10% more accuracy and 20% more specificity than those of the baseline models). Experimental results also show that eyes are the most affected facial landmark due to dehydration. Our smartphone-based dehydration diagnosis tool may not achieve the accuracy level of the clinical diagnosis by professional medical equipment or expert personnel. However, in most cases, our solution can provide crucial insights and prompt the users to take necessary actions early on. We aim to extend our solution to detect different dehydration levels (e.g., mild, moderate, severe) in the future.

Acknowledgements. This research has been done at Bangladesh University of Engineering and Technology (BUET). We thank all the volunteers who participated with their valuable data to make this research possible.

References

1. Ahmed, S.M., Hossain, M.A., RajaChowdhury, A.M., Bhuiya, A.U.: The health workforce crisis in Bangladesh: shortage, inappropriate skill-mix and inequitable distribution. Hum. Resour. Health **9**(1) (2011). https://doi.org/10.1186/1478-4491-9-3
2. Armstrong, L.E., et al.: Mild dehydration affects mood in healthy young women. J. Nutr. **142**(2), 382–388 (2012)
3. Barrell, A.: What to know about dehydrated skin (2021). https://www.medicalnewstoday.com/articles/dehydrated-skin#causes. Accessed 14 Feb 2022
4. Bilal, S., et al.: Evaluation of standard and mobile health-supported clinical diagnostic tools for assessing dehydration in patients with diarrhea in rural Bangladesh. Am. J. Trop. Med. Hyg. **99**(1), 171–179 (2018)
5. Chicco, D.: Siamese neural networks: an overview. In: Cartwright, H. (ed.) Artificial Neural Networks. MMB, vol. 2190, pp. 73–94. Springer, New York (2021). https://doi.org/10.1007/978-1-0716-0826-5_3
6. Chollet, F.: Xception: deep learning with depthwise separable convolutions, pp. 1800–1807 (2017). https://doi.org/10.1109/CVPR.2017.195
7. Dehydrated? How not drinking enough water impacts your eyes (n.d.). https://www.essilorusa.com/newsroom/dehydrated-how-notdrinking-enough-water-impacts-your-eyes. Accessed 14 Feb 2022
8. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
9. Dry lips (n.d.). https://www.healthline.com/health/dehydration-white-tongue#other-symptoms. Accessed 29 Apr 2022
10. Fukushima, Y., et al.: A pilot clinical evaluation of oral mucosal dryness in dehydrated patients using a moisture-checking device. Clin. Exp. Dent. Res. **5**(2), 116–120 (2019)

11. Gairola, S., et al.: SmartKC: smartphone-based corneal topographer for keratoconus detection. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **5**(4) (2022)
12. Ganio, M.S., et al.: Mild dehydration impairs cognitive performance and mood of men. Br. J. Nutr. **106**(10), 1535–1543 (2011). https://doi.org/10.1017/s0007114511002005
13. Ghojogh, B., Sikaroudi, M., Shafiei, S., Tizhoosh, H.R., Karray, F., Crowley, M.: Fisher discriminant triplet and contrastive losses for training Siamese networks. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE (2020). https://doi.org/10.1109/ijcnn48605.2020.9206833
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
15. Koch, G.R.: Siamese neural networks for one-shot image recognition (2015)
16. Leiper, J.B., Molla, A.M.: Effects on health of fluid restriction during fasting in Ramadan. Eur. J. Clin. Nutr. **57**(S2), S30–S38 (2003). https://doi.org/10.1038/sj.ejcn.1601899
17. Levine, A.C., et al.: External validation of the DHAKA score and comparison with the current IMCI algorithm for the assessment of dehydration in children with diarrhoea: a prospective cohort study. Lancet Glob. Health **4**(10), e744–e751 (2016). https://doi.org/10.1016/s2214-109x(16)30150-4
18. Liaqat, S., Dashtipour, K., Arshad, K., Ramzan, N.: Non invasive skin hydration level detection using machine learning. Electronics (Basel) **9**(7), 1086 (2020)
19. Liu, C., Tsow, F., Shao, D., Yang, Y., Iriya, R., Tao, N.: Skin mechanical properties and hydration measured with mobile phone camera. IEEE Sens. J. **16**(4), 924–930 (2016)
20. Liu, G., Smith, K., Kaya, T.: Implementation of a microfluidic conductivity sensor—a potential sweat electrolyte sensing system for dehydration detection. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, pp. 1678–1681. IEEE (2014). https://doi.org/10.1109/EMBC.2014.6943929
21. Mariakakis, A., Banks, M.A., Phillipi, L., Yu, L., Taylor, J., Patel, S.N.: Biliscreen: smartphone-based scleral jaundice monitoring for liver and pancreatic disorders. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(2) (2017). https://doi.org/10.1145/3090085
22. Mariakakis, A., et al.: Pupilscreen: using smartphones to assess traumatic brain injury. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(3), 81 (2017). https://doi.org/10.1145/3131896
23. Ozana, N., et al.: Improved noncontact optical sensor for detection of glucose concentration and indication of dehydration level. Biomed. Opt. Express **5**(6), 1926–1940 (2014)
24. Perkins, B.A., et al.: Evaluation of an algorithm for integrated management of childhood illness in an area of Kenya with high malaria transmission. Bull. World Health Organ. **75**(Suppl. 1), 33–42 (1997)
25. Reljin, N., et al.: Automatic detection of dehydration using support vector machines. In: 2018 14th Symposium on Neural Networks and Applications (NEUREL), Belgrade. IEEE (2018)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
27. Skin turgor (n.d.). https://medlineplus.gov/ency/article/003281.htm. Accessed 14 Feb 2022

28. Sunken eyes (n.d.). https://www.healthline.com/health/dry-eye/ask-the-expert-dry-eye-dehydration#hydrating. Accessed 29 Apr 2022
29. Wikipedia contributors. Firebase—Wikipedia, the free encyclopedia (2021). https://en.wikipedia.org/w/index.php?title=Firebase&oldid=1054631697. Accessed 15 Feb 2022
30. Wikipedia contributors. Flutter (software)—Wikipedia, the free encyclopedia (2022). https://en.wikipedia.org/w/index.php?title=Flutter_(software)&oldid=1070411100. Accessed 15 Feb 2022
31. Wikipedia contributors. Siamese neural network—Wikipedia, the free encyclopedia (2021). https://en.wikipedia.org/w/index.php?title=Siamese_neural_network&oldid=1020522415. Accessed 15 Feb 2022
32. Xu, X., et al.: Listen2cough: leveraging end-to-end deep learning cough detection model to enhance lung health assessment using passively sensed audio. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **5**(1), 42:1–42:22 (2021). https://doi.org/10.1145/3448124
33. Zdolsek, J., Li, Y., Hahn, R.G.: Detection of dehydration by using volume kinetics. Anesth. Analg. **115**(4), 814–822 (2012)

Software Frameworks and Interoperability

Experiencer: An Open-Source Context-Sensitive Wearable Experience Sampling Tool

Alireza Khanshan[1](✉), Pieter Van Gorp[2], and Panos Markopoulos[1]

[1] Department of Industrial Design, Eindhoven Artificial Intelligence Systems Institute, Eindhoven University of Technology,
5612 Eindhoven, AZ, The Netherlands
`{a.khanshan,p.markopoulos}@tue.nl`

[2] Department of Industrial Engineering, Eindhoven University of Technology, 5612 Eindhoven, AZ, The Netherlands
`p.m.e.v.gorp@tue.nl`

Abstract. We introduce Experiencer, a newly developed Experience Sampling Method (ESM) software for commodity-level smartwatches. We designed this software mainly to address the compliance-related challenges, such as dropouts of study participants, that generations of ESM software solutions have faced. Dropouts are often caused by the inconvenient frequency and timing of the ESM prompts. This can partly be mitigated by utilizing physiological smartwatch sensors to learn which prompting moments are both convenient to the study participant and also relevant to the ESM study designer. Experiencer enables researchers to configure context-sensitive sampling protocols, providing access to raw sensor data, within the boundaries of European privacy legislation. In this paper, we describe the technical capabilities of our software, compare its features with the state-of-the-art, and showcase its application in studies that used Experiencer.

Keywords: Experience Sampling Method · Wearable ESM · mHealth · Ubiquitous Computing · Smartwatch application · Wearables · Software Framework

1 Introduction

Studying what people do, feel, and think during their daily routines is essential to understanding the dynamics of the human psyche [13]. Although the interest for such information originates in psychological research, it extends to a variety of other domains where humans are studied (e.g., healthcare [33], media studies [43], or education [6]). Actions and thoughts can be registered in diaries. Registering frequently and in a real-life context can result in entries with high ecological validity [41]. However, poor compliance to journaling protocols and human forgetfulness are well-documented challenges for the diary method [10]. Therefore, it is preferable to distribute self-report prompts throughout daily life rather

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 315–331, 2023.
https://doi.org/10.1007/978-3-031-34586-9_21

than journaling retrospectively [3,32]. Such a process of collecting self-reported data about behaviors, thoughts, or feelings during the day-to-day activities of humans, is commonly called the Experience Sampling Method (ESM [32]). The same approach with an emphasis on psychological research is known as Ecological Momentary Assessment (EMA [48]). In the field of human-computer interaction, context sensing technologies were emerging two decades ago and led to the reintroduction of ESM as Contextual Aware Experience Sampling (CAES [24]) with the emphasis on utilizing context sensing to optimize the sampling procedures. For the sake of simplicity and consistency, we use the term ESM in the remainder of this paper to encompass the aforementioned concepts. Additionally, we introduce the term wESM (wearable ESM) to refer to the use of wearables (e.g., smartwatches) instead of or together with smartphones or other mediums to handle both collection of self-reported data as well as sensor data.

Considering that ESM aims to prompt during daily life, finding opportune moments that do not interfere with one's daily activities to deliver such self-report prompts (beeps) becomes essential. By doing so, the likelihood of disturbing study participants becomes lower, hence, compliance is potentially increased [29]. One common approach to detecting opportune moments is by longitudinal monitoring via physiological sensors to better perceive the momentary context of respondents.

In the context of wESM, interpretation of sensor data usually requires some knowledge about the context in which the sensor was used as well as the participant-specific subjective data. Thus, self-reports are critical to provide ground truth and complement the sensor data. Then again, study participants may skip self-reports, especially when studies last longer than a few days. How to minimize dropouts and maximize compliance requires further research, especially in the context of longitudinal studies. Typically, compliance is measured by calculating the dropout rate, response rate, response time, resolution time, volunteering rate, and the amount of presented information which oftentimes indicated the poor compliance of the participants [18,31,35,41,50,54].

Arguably, the choice of ESM device used for prompting and data entry, as well as the prompting schedule affects the participation experience and consequently impacts the extent to which the ESM process is perceived as seamless or in contrast, burdensome by study participants [20,23,29,34,49]. Recent advances in commodity-level smartwatches in terms of interactivity, connectivity, and embedded sensing technology, offer new opportunities for using them as wESM devices, which could help reduce the obtrusiveness of wESM signals and increase their availability as they are wrist-worn. Furthermore, they provide a quick-and-concise data entry interface, and their sensors pave the way to provide context-sensitive prompting regimes [46,47]. In the following, we introduce Experiencer [26], our open-source context-sensitive smartwatch-based wESM software that provides researchers with advanced data gathering features while striving to attain compliance of participants. Experiencer has been designed with the aim to enable researchers to flexibly configure their experiment protocol to potentially alleviate response fatigue and sustain sufficient response rates that are otherwise hampered by traditional scheduling regimes (e.g., random sampling) during ESM studies [12,18,55].

In the following, we first review research on ESM-related devices to derive requirements for wESM support. Subsequently, we introduce Experiencer by providing conceptual and technical information on client and server components and their interactions. We also demonstrate how Experiencer was already used in a variety of wESM studies. Afterward, we discuss limitations, weaknesses, and directions for future work. Finally, we summarize our contributions in the conclusion section and encourage the readers to utilize Experiencer for their own research.

2 Related Works

In this section, we initially summarize the history of digital ESM devices and their evolution over time. Then, we focus on the contemporary smartwatch-based wESM solutions and examine the state-of-the-art in the domain.

2.1 History of Digital ESM Devices

While the experience sampling *method* emerged already in the eighties [32], digital devices were gaining popularity around the turn of the century (e.g., Electronic Mood Device [22] by Hoeksma et al. or the Experience Sampling Program [3] by Barret et al.).

Due to the eventual prevalence of smartphones since the late 2000s,s, and the software development kits (SDK) supported by their operating systems (early on by PalmOS and Windows CE, and more recently by Android, and iOS), more advanced ESM tools have been created. The applications developed in the context of ESM have mainly sought to alleviate the complexity of configuring ESM protocols for researchers by offering custom configuration schemes (e.g., Momento [11], AndWellness [21], PsyMate [36], Tempest [4], [5], and formR [2]) and facilitating research-focused data collection (e.g., Funf [1], RADAR-base [42], and HOPES [53]). Meanwhile, fewer works have accounted for context awareness. By incorporating context-sensitive strategies the researchers can capture data at specific events (e.g., context-aware experience sampling tool [24], AWARE [17], and Paco [16]).

More recently, commodity-level wearables such as smartwatches are being utilized in the context of ESM as well. They offer SDKs on par with smartphones and their physiological sensors are more accurate and reliable since they are on the skin, rather than in the pocket. In addition, they are optimized to collect physiological (e.g., body activity) data more continuously.

2.2 Towards wESM Smartwatch Applications

In this section, we survey the state-of-the-art in the application domain for smartwatches utilized for ESM studies.

Intille et al. focused on the amount and length of interruption, and the difficulty of accessing the device of typical Ecological Momentary Assessment (EMA) delivered via smartphones [23]. They implemented μEMA as a smartwatch extension to smartphones that delivered prompts on the smartwatch as well as concise versions of ESM questions. A study was conducted where μEMA on a smartwatch was compared with EMA exclusively on a phone. Despite an $\approx$8 times increase in the number of interruptions, μEMA had a significantly higher compliance rate, completion rate, and first prompt response rate, and was perceived as less distracting. Although μEMA [23] suggested that a substantially higher prompting rate than EMA, may yield higher response rates and a lower participation burden, Ponnada et al. aimed to assess the validity of participant responses from μEMA self-reports [40]. It was concluded that for physical activity registrations, high-frequency μEMA self-reports were consistent with activities detected by a research-grade continuous sensor, even when prompting up to 72 times per day. This demonstrated that μEMA study participants were not carelessly answering prompts by randomly tapping on the smartwatch. Then again, the experiment by Ponnada et al. lasted only one week, so further development of smart prompting protocols may be needed to reduce participant burden and enhance compliance in longer studies.

Blaauw et al. presented Physiqual, a platform for researchers that gathers and integrates data from commercially available sensors and service providers into one unified format for use in ESM, and Quantified Self (QS) [9]. The Physiqual platform allows researchers to aggregate and integrate physiological sensor data with ESM. Although such a platform does not provide a dedicated wESM application for smartwatches, it facilitates the aggregation of such pre-existing data sources.

Kheirkhahan et al. developed a smartwatch-based framework for real-time and online assessment and mobility monitoring (ROAMM) [30]. The smartwatch application was used to collect and pre-process data. A server was used for storing and retrieving data, remote monitoring, data visualization, and summary statistics, and for other administrative purposes. Although the smartwatch app allowed configurable sensors and supports different types of studies, it is not openly available and does not support context-sensitive scheduling.

Hafiz et al. showed a strong correlation between the data gathered via their domain-specific smartwatch application and computer-based tests in a lab setting [19]. The aims of their study were to evaluate the Ubiquitous Cognitive Assessment Tool (UbiCAT), a smartwatch-based platform they developed to assess cognitive performance, to investigate its usability, and to understand participants' perceptions regarding the use of a smartwatch in cognitive assessment.

Park et al. developed a framework for collecting and analyzing physiological data using smartwatches in the wild and demonstrated its robustness away from controlled laboratory settings [38]. Their system sent random notifications during the day asking questions about subjective well-being. They concluded that methodological research needs to study how to interpret continuous physiological signals obtained through such platforms. Once such understanding is developed, sensing can potentially reduce the burden of self-reporting.

Collectively, these works demonstrate the feasibility, benefits, and pitfalls of using smartwatches in ESM studies. Unfortunately, however, the underlying software systems are not available for elaboration by other scholars.

3 Requirements for a Smartwatch-Based wESM Tool

In our analyses of the aforementioned state-of-the-art, we observed that none of the platforms was offered as a free, reliable, and continued service to other academics. Furthermore, we identified the lack of a thorough exploration of the design space concerning commodity-level smartwatches. In the following, we propose our software requirements for designing a smartwatch-based wESM tool that we realized by analyzing existing solutions, exploring the design space, and empirically during the development of Experiencer.

The flexible nature of ESM protocols demands guidelines that researchers can follow to set up their ESM studies. However, earlier considerations on tools to support ESM studies(e.g., [14], [44], and [39]) need to be reviewed to address the opportunities and challenges introduced by wearable devices and modern software technology. For example, the non-reactivity guideline proposed by Delespaul [14]: reactivity can be minimized by using small, reliable, and inexpensive devices that produce unpredictable prompts and can be fully employed within a range of environmental constraints [14]. Such concerns can now be addressed by using a commodity-level smartwatch that can run configurable wESM software.

By surveying the latest developments in this domain, literature study, running our own experiments each with at least 50 participants (the number of participants was determined based on power analysis, the number of available smartwatches, and the recruitment process [29,37]), and discussing the requirements with respective ESM researchers, we distilled a shortlist of features that wESM platforms should provide. We found four highly relevant multi-purpose wESM solutions that explicitly used the smartwatches as their main client or specifically developed to incorporate smartwatches as third-party devices in their ecosystem. We realized six categories of features to list the key similarities and differences between the aforementioned solutions: *Data collection and analysis*, *Scheduling*, *Data entry*, *Monitoring interface*, *Scalability*, and *Optimization*. Moreover, we added *Openness, availability, and security* as an additional category, to make clear that no solution so far could be used by other scholars continuously.

Table 1 provides a comparative overview of the software features of the aforementioned platforms in those seven categories. We define these features as follows:

Recording sensor data refers to the recording and storing of the data captured via physiological sensors such as an accelerometer and a heart rate monitor.

Table 1. Feature Comparison of recent wESM platforms

Category	Feature	Software				
		WellBeat [38]	*ROAMM* [30]	*μEMA* [23]	*Physiqual* [9]	*Experiencer* [26]
Data collection and analysis	Recording sensor data	✓	✓	✓	✓	✓
	Data analytics dashboard	✓	✓	✗	✓	✗
	Configurable sensors	✗	✓	✗	✗	✓
Scheduling	Context-sensitive	✗	✗	✗	✗	✓
	Temporal	✗	✓	✗	✗	✓
Data entry	Input widgets on the smartwatch	✓	✓	✓	✗	✓
	Configurable widgets	✗	✓	✗	✓	✓
Web interface	Data visualization	✗	✓	✓	✗	✗
	Administration dashboard	✗	✓	✗	✓	✗
Scalability	Remote device management	✗	✓	✗	✗	✓
Optimization	Event-based data collection	✗	✗	✗	✗	✓
	Custom data synchronization	✓	✓	✓	✗	✓
Openness, availability, and security	Reusability and availability	✗	✗	✗	✗	✓
	GDPR compliance	✗	✗	✗	?	✓

Data analytics dashboard enables the realization of statistical reports and/or discovery of patterns in the collected raw data after data is stored on the server. Such processing can ideally be customized by the study owner.

Configurable sensors indicates the possibility to enable the desired set of sensors, setting specific sampling frequencies, and setting the duration of the sensor data recording period.

Context-sensitive refers to the possibility of sending beeps at contextual events of interest rather than at random or interval-based times. Such events can refer to simpler conditions such as "send a prompt *when sedentary activity is detected*", or more complex ones such as "send a prompt *5 min after a running activity is ended*", or "send a prompt *as soon as the minimum heart rate during 10 straight minutes is higher than 100bmp*".

Temporal refers to the capability of defining wESM protocols, through configurable settings rather than hard-coded, in time-based manners that are signal-contingent (i.e., random) or interval-contingent (as defined in [8,14]).

Input widgets on the smartwatch turn the wearable into a data entry device rather than just a notification device (which would still require smartphone interactions). Typical widgets are *text inputs*, *radio buttons*, or *drop-downs.*

Configurable widgets facilitates the creation or customization of the look and feel of input widgets. This is usually supported by either HTML scripting or by parameterizing pre-built widget components.

Data visualization provides computer-generated representations of the data in form of charts (e.g., histogram, and scatter plot) in a dedicated dashboard.

Administration dashboard allows the creation and management of wESM protocols, monitoring the activity of participants, and accessing the collected data of the study via a graphical user interface (GUI).

Remote device management allows controlling and monitoring the wESM electronic device (e.g., smartwatch) remotely. Such a feature allows fast and easy scale-out by facilitating the configuration of numerous devices at once. The control over the device also helps with restraining specific out-of-the-box features of the device (e.g., disabling GPS to enhance battery life and ensure privacy, disallowing the study participants to install apps on their device, or preventing factory reset). Remote access can also help with updating the aforementioned constraints on the fly and seamlessly during the study according to researchers' requirements or participants' convenience.

Event-based data collection enables intermittent collection of data triggered by events of interest; e.g., recording heart rate data solely during answering a questionnaire. This eases the matching of sensory data with self-report periods and greatly reduces battery consumption compared to when data is recorded continuously.

Custom data synchronization supports the buffering of data on the wearable and smart synchronization with the server, potentially also balancing battery life with the information needs of the study owner. The wESM app may monitor the WiFi coverage, assess the Internet connection stability, and then transfer the data in controlled transactions to assure data persistence and consistency.

Reusability and availability is unfortunately seldom seen in the latest wESM solutions. Most of them have either become obsolete already (not maintained after a specific study), or they were never designed to be reusable (only for a specific use case, research question, or domain). The few solutions that are potentially reusable were never provided as a service to other scholars, Experiencer is open-source [27] and its back-end, GameBus, is open-access [51].

GDPR compliance The General Data Protection Regulation (GDPR) is a legal framework that sets guidelines for the collection and processing of personal information from individuals who live in the European Union (EU) [15]. Compliance with the GDPR would lessen privacy concerns especially for EU users (be it researchers or participants). During our assessment, we could not find relevant or explicit information regarding the GDPR and privacy policy for most of the solutions. In the case of Physiqual, a platform that connects the different third-party tools, its conformity partially relies on such third-party tools (e.g., Google Fit, Fitbit, etc.). Thus, we put **?** in the table.

Note that none of the four tools support context-sensitive sampling. Nonetheless, such a feature can contribute greatly to the study participants (i.e., interruptions or prompts are received at more opportune moments resulting in less burden and fatigue), as well as the researchers (i.e., responses recorded at more opportune moments are less biased). Therefore, we did treat it as a requirement for Experiencer. The same holds for the event-based data collection feature.

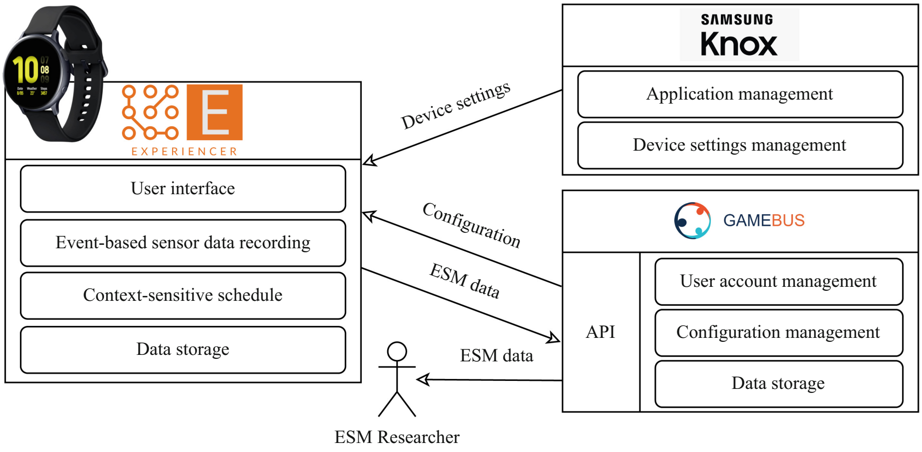


Fig. 1. Overview of system components and communications

4 Configuring wESM Protocols with Experiencer

Experiencer is a context-sensitive wESM tool that allows for recording of sensor data, configuring sensors, remote device management, event-based data collection, various sampling regimes, dynamic user interface (UI), and also optimizes device data storage and data transactions over the network while being open-access, available, and compliant with standard privacy measures. The back-end of Experiencer is built on top of GameBus, an open-access health data management platform that is offered non-commercially by academia [45]. GameBus, designed following the GDPR-oriented privacy and security measures, guarantees that all data is stored exclusively in Europe and provides to its users full control over their data. Experiencer also builds upon Knox, an industrial-strength device management system by Samsung. Specifically, Knox is used for the remote (re-)installation and (re-)configuration of Experiencer. Figure 1 visualizes this modular software architecture.

The behavior of Experiencer is defined by its *configuration* [27]. The current version of Experiencer allows setting the configuration through its API. Composing a configuration is the first step to conduct a study using Experiencer. That includes setting an inter-notification time value, specifying sensor(s) settings, defining contextual rules for context-sensitive sampling (if not, the fixed interval policy is adopted automatically), and defining a questionnaire. Table 2 overviews the capabilities proposed by our configurations. Each participant's account is linked to a specific smartwatch. Next, the participant-smartwatch pairs are linked to a configuration and a study (that is named by the researcher) (Fig. 2). During the linkage, the researcher can apply different configurations to the participants of the same study. By setting the configuration during the linkage procedure, the researcher can construct their treatment groups (e.g., by setting some participants to configuration A and some others to configuration

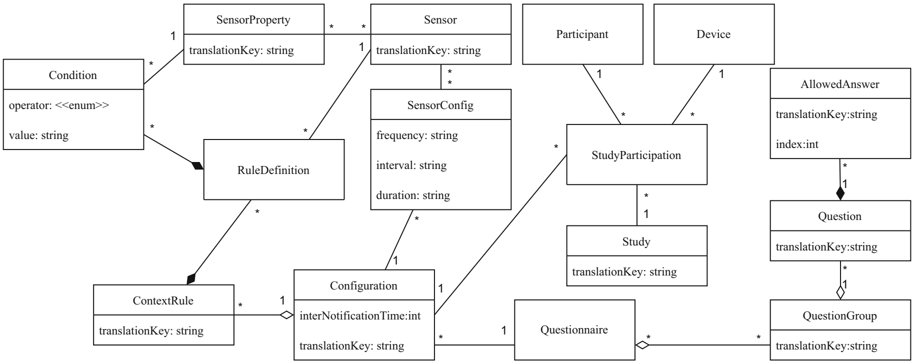


Fig. 2. Provisional UML class diagram of Participant, Device, Configuration, and Questionnaire. All translationKey attribute values should be unique.

B). Lastly, the final configuration (including the inter-notification time, sensor settings, questionnaire, etc.) is transferred from the server to the smartwatch over WiFi. After the successful transmission, the app is ready to use. In addition, the app periodically checks for configuration updates (e.g., the addition of a new question to the questionnaire, or a new value for the inter-notification time), so changes can be applied even after devices have been deployed to study participants.

Experiencer consumes a JSON-formatted configuration [27]. An essential part of this configuration is its questionnaire (Fig. 2). Experiencer is equipped with a parser that interprets the questionnaire dynamically and renders a layout based on the configurations. The current version of Experiencer supports questions that are chained in a sequence (e.g., a 2-question PANAS questionnaire starts with Fig. 3a followed by Fig. 3b). Additionally, if a researcher desires a specific and more complex UI for the questionnaire, they can program their own interface using HTML, CSS, and JavaScript.

Questionnaires in Experiencer are made up of question groups where each holds a set of one or more questions with their corresponding answers (Fig. 2). Question groups allow the researchers to categorize their questions. For example, a questionnaire can start with a number of *general* questions followed by a set of *personality*-related questions (*general* and *personality* are example categories). Such grouping aims at a more distinctive and informative user experience.

The key notion in creating a questionnaire is the *translation keys*. Translation keys are unique strings that serve both as identifiers and localization means. They are used to look up the GameBus-provided human-readable texts (translations) in different languages. Given such translations, the questionnaire object of the configuration can be transformed into a UI (similar to Fig. 3a and Fig. 3b). The generated UI is shown to the participants when they press the self-report button (Fig. 3c), or it appears automatically after a scheduled prompt. Once a participant responds to a question, their response is stored locally and then transferred to the server transactionally when a stable network connec-

Table 2. Configuration options of Experiencer

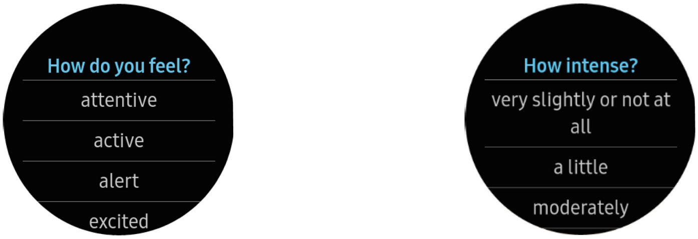
Configuration	Options	Description
Sampling policy	Context-sensitive sampling or fixed interval	Experiencer is programmed to read data from different sensor and can be configured to beep based on sensor data. For example, Experiencer can be configured to send beeps only when study participants engage in vigorous physical activities or vice versa [29]. The fixed interval policy provides the classic interval-based prompting regime when desired.
Inter-notification time	An integer value indicating inter-notification time	The inter-notification time, determines the time in-between each beep. The role of inter-notification time for the fixed interval policy is to determine the period between two beeps. For context-sensitive policies, the inter-notification time determines the cool-down period.
Unobtrusive sensing	Accelerometer, photoplethysmography, heart rate, peak-peak interval, body activity sensors	Body activity data is continuously monitored to ensure the accuracy of event-contingent policies. It also serves as a means to know if the smartwatch is being worn. The other sensors, if chosen by the researcher, are recorded during the period that a participant is filling an ESM form (e.g., a questionnaire), by default, for a maximum of 1 min. Continuous recording is also possible.
Compliance status	Timestamps related to beep received, read, and response submission times	To analyze compliance, the time when a beep is received, read, and submitted are by default recorded to facilitate the calculation of compliance-related indices such as response delay.
Questionnaire	Questions and answers as string literals	The set of ESM questions can be defined by the researcher. The list is then parsed within the app and represented to the study participant when a self-report procedure is started. The questionnaire in the current version is sequential rather than branched [14]

tion is detected. The GameBus back-end already pre-defines a variety of such questions. Still, researchers with ESM protocols involving custom questions can request the addition of such items [52].

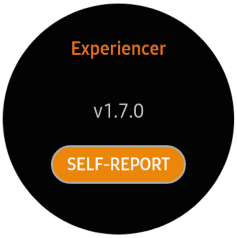
5 Case Studies Run with Experiencer

We created Experiencer for 1) effective and accurate context sensitivity to help increase compliance, and 2) openness, availability, and security. Regarding the former, not only do we have run various wESM studies but also continuously plan to test different hypotheses regarding context percipiency. Our learned lessons from such studies help us fine-tune the context-sensitivity of Experiencer over time by continuously training models that can be leveraged for decision-support on expected compliance of protocols. For the latter, on the one hand, we allow interested researchers from various domains to set up their experiments with Experiencer. Moreover, we actively maintain and develop our software following the best practices and guidelines of software engineering and security [26].

To demonstrate the aforementioned capabilities and flexibility of our solution, below we describe the different wESM studies that utilized Experiencer along with their goal, and their outcome. The study-specific configuration files are publicly accessible via the project's GitHub examples repository [27].



(a) PANAS question (1st in sequence) (b) PANAS question (2nd in sequence)



(c) The main screen of Experiencer

Fig. 3. Screenshots of Experiencer application as it was configured for the SamenGezond 2020 and 2021 campaigns

The SamenGezond 2020 and 2021 Campaigns. During two health promotion campaigns [37] we used Experiencer, to assess the effects of physical activity upon experience sampling response rate on smartwatches. We adopted a context-sensitive schedule to prompt half of the participants at subtle and the other half at vigorous physical activity levels and observed a significant difference in response rates depending on such context [29].

GGz Centraal. To measure and predict the stress levels of a subgroup of GGz Centraal mental healthcare facility patients, Experiencer is being used since 2021 to collect valence and arousal data throughout the day. This ongoing study also helps assess the adaptability of Experiencer in targeting various cohorts. To meet the requirements of the target group, a custom user interface was created by exploiting the rotary capabilities of the smartwatches.

Persuasion Profiling. In another wESM study, Experiencer was used to capture student motivation, and to assess its influence on the response rate of students while applying persuasion profiling. Persuasion profiling involves tailoring notifications to users (the content of experience sampling prompts in this case) according to an individual's susceptibility to known social influence strategies [25]. On the other hand, the researcher's script to generate a custom data entry widget was incorporated.

Affective State. To complement our findings concerning opportune moments of interruption based on the context [29], in a collaborative wESM study to measure time-lagged associations between affective state, sleep, and several other lifestyle-related behaviors, Experiencer was set up to deliver experiential questions compliant with a context-sensitive schedule based on physical activity. Following the researcher's constraints, the app was configured to send beeps between habitual wake-up and sleep times.

6 Future Work and Limitations

We have sought to identify the requirements of researchers and study participants so far. Markedly, the addition of branching questionnaires where the transition of one question to the next is controlled by a rule-based system is crucial. Such a rule-based system takes into account the response to a question and/or perceived context to determine the subsequent question(s). Additionally, the context-sensing could be improved by including contextual information derived from device usage patterns or others alike related to participants' behavior to complement the sensory rules. Moreover, a smartwatch-based wESM software should include a wide range of user interface elements such as checkboxes, radio buttons, sliders, etc. to support a substantial subset of ESM studies (excluding those that are currently difficult to do because of the small smartwatch screen size and difficulty of text input). In addition, to simplify the interactions between the researcher and wESM software, interactive dashboards for the administration of wESM protocol configuration, as well as real-time monitoring and visualization of the data should be designed. Indeed such flexibilities, by design, do not guarantee higher compliance with respect to response rate and retention since researchers may apply sampling regimes that are deemed intrusive. Accordingly, the domain lacks fine-grained ESM-related data that would otherwise enable data-intensive approaches such as data-driven modeling and machine learning to help address compliance-related challenges in the ESM domain especially by learning response patterns [28,56]. Thus, more collective effort is required to share such data openly with other scholars.

Regardless, smartwatch-based wESM tools are affected by some limitations, such as their rather short battery life especially when sensor data is recorded for longer periods of time and with high frequencies. Although Experiencer provides enough flexibility to alter the period of the data collection and the sampling frequency of the sensors based on the researcher's requirements, there is an inevitable trade-off between the accuracy (and volume) of the collected sensor data and the battery life for wESMs.

7 Conclusions

Experience Sampling is a very popular research method that is used in multiple fields to study the experiences and thoughts of respondents over sustained periods of time, by repeatedly prompting them to respond to survey questions. Experience sampling presents several methodological challenges, like participant burden, low compliance, etc., which researchers have traditionally attempted to address through technological developments. While currently, most experience sampling studies rely on smartphones [7], the form factor and sensing capabilities of smartwatches and their growing prevalence open up new opportunities for researchers wishing to apply experience sampling studies. This paper reviewed the few attempts that have been made so far to support ESM studies with wearable devices and discussed the requirements for modern ESM tools that exploit the recent technological advancements in smartwatches. We described the state of the art in the emerging area of wESM, where wearables and more specifically smartwatches can be used as signaling and reporting devices in ESM studies. We created Experiencer; a context-sensitive wearable experience sampling application on commodity-level smartwatches. By perceiving context through physiological sensors, our software allows researchers to configure sampling regimes that potentially align with the opportune moments of interrupts. Thus, the likelihood of sustaining sufficient compliance is increased. Besides that, the input widgets are all on the smartwatch, accessible on their wrist, and are resolvable with just a few taps. Besides, context-sensitiveness helps with interrupting at more opportune moments, thus, the participants are potentially less likely to be disturbed. We expect that such capabilities result in less burden for the participants during the study period, and improved compliance and data quality. Future research could evaluate through case studies the extent to which these benefits can be delivered. Furthermore, we argued for a specific set of requirements that wESM platforms need to address, and have shown how Experiencer improves them compared to the state-of-the-art and emphasized how Experiencer advances software availability and context sensitivity. Moreover, we describe the technical aspects of our software, look into its configurable features and review how Experiencer was configured and used in different ESM studies. Lastly, we acknowledge the improvements required in this domain and point out the barriers caused by the device restrictions and constraints.

Acknowledgment. Special recognition to Jory Schoondermark and Barbara Montagne at GGz Centraal, as well as Karin Smolders and Lars Giling at the Eindhoven University of Technology who have contributed to the development of Experiencer through their input and feedback on early versions of Experiencer. Lastly, special appreciation goes to Samsung Nederland for their excellent support on the smartwatches we utilized for our research and development.

References

1. Aharony, N., Gardner, A., Sumter, C.: funf | open sensing framework (2021). http://www.funf.org/
2. Arslan, R.C., Walther, M.P., Tata, C.S.: formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. Behav. Res. Methods **52**(1), 376–387 (2020). https://doi.org/10.3758/s13428-019-01236-y
3. Barrett, L.F., Barrett, D.J.: An introduction to computerized experience sampling in psychology. Soc. Sci. Comput. Rev. **19**(2), 175–185 (2001). https://doi.org/10.1177/089443930101900204
4. Batalas, N., Markopoulos, P.: Introducing tempest, a modular platform for in situ data collection. In: Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design, pp. 781–782 (2012)

5. Batalas, N., aan het Rot, M., Khan, V.J., Markopoulos, P.: Using tempest: End-user programming of web-based ecological momentary assessment protocols. Proceedings ACM Hum. Comput. Interact. **2**(EICS), 1–24 (2018)
6. Becker, E.S., Goetz, T., Morger, V., Ranellucci, J.: The importance of teachers' emotions and instructional behavior for their students' emotions-an experience sampling analysis. Teach. Teach. Educ. **43**, 15–26 (2014)
7. van Berkel, N., Ferreira, D., Kostakos, V.: The experience sampling method on mobile devices. ACM Comput. Surv. (CSUR) **50**(6), 93:1–93:40 (2017). https://doi.org/10.1145/3123988
8. van Berkel, N., Goncalves, J., Lovén, L., Ferreira, D., Hosio, S., Kostakos, V.: Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. Int. J. Hum Comput Stud. **125**, 118–128 (2019). https://doi.org/10.1016/j.ijhcs.2018.12.002
9. Blaauw, F.J., et al.: Let's get physiqual - an intuitive and generic method to combine sensor technology with ecological momentary assessments. J. Biomed. Inf. **63**, 141–149 (2016). https://doi.org/10.1016/j.jbi.2016.08.001
10. Bolger, N., Davis, A., Rafaeli, E.: Diary methods: capturing life as it is lived. Annu. Rev. Psychol. **54**(1), 579–616 (2003)
11. Carter, S., Mankoff, J., Heer, J.: Momento: support for situated ubicomp experimentation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 125–134. Association for Computing Machinery (2007). https://doi.org/10.1145/1240624.1240644
12. Collins, R.L., Kashdan, T.B., Gollnisch, G.: The feasibility of using cellular phones to collect ecological momentary assessment data: application to alcohol consumption. Exp. Clin. Psychopharmacol. **11**(1), 73–78 (2003). https://doi.org/10.1037/1064-1297.11.1.73
13. Csikszentmihalyi, M., Larson, R.: Validity and reliability of the experience-sampling method. In: Flow and the Foundations of Positive Psychology, pp. 35–54. Springer, Dordrecht (2014). https://doi.org/10.1007/978-94-017-9088-8_3
14. Delespaul, P.A.E.G.: Technical note: devices and time-sampling procedures. In: Vries, M.W.d. (ed.) The Experience of Psychopathology: Investigating Mental Disorders in their Natural Settings, pp. 363–374. Cambridge University Press (1992). https://doi.org/10.1017/CBO9780511663246.033
15. European Parliament: General Data Protection Regulation (GDPR) (2016). https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04
16. Evans, B.: Paco-applying computational methods to scale qualitative methods, vol. 2016, no. 1, pp. 348–368 (2016). https://doi.org/10.1111/1559-8918.2016.01095
17. Ferreira, D., Kostakos, V., Dey, A.K.: AWARE: mobile context instrumentation framework. Front. ICT **2**, 6 (2015). https://doi.org/10.3389/fict.2015.00006
18. Fuller-Tyszkiewicz, M., Skouteris, H., Richardson, B., Blore, J., Holmes, M., Mills, J.: Does the burden of the experience sampling method undermine data quality in state body image research? Body Image **10**(4), 607–613 (2013). https://doi.org/10.1016/j.bodyim.2013.06.003
19. Hafiz, P., Bardram, J.E.: The ubiquitous cognitive assessment tool for smartwatches: design, implementation, and evaluation study. JMIR Mhealth Uhealth **8**(6), e17506 (2020). https://doi.org/10.2196/17506
20. Hernandez, J., McDuff, D., Infante, C., Maes, P., Quigley, K., Picard, R.: Wearable ESM: differences in the experience sampling method across wearable devices. In: Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 195–205. MobileHCI 2016, Association for Computing Machinery (2016). https://doi.org/10.1145/2935334.2935340

21. Hicks, J., Ramanathan, N., Kim, D., Monibi, M., Selsky, J., Hansen, M., Estrin, D.: AndWellness: an open mobile system for activity and experience sampling. In: Wireless Health 2010, pp. 34–43. WH 2010, Association for Computing Machinery (2010). https://doi.org/10.1145/1921081.1921087
22. Hoeksma, J.B., Sep, S.M., Vester, F.C., Groot, P.F.C., Sijmons, R., De Vries, J.: The electronic mood device: design, construction, and application. Behav. Res. Methods Instrum. Comput. **32**(2), 322–326 (2000). https://doi.org/10.3758/BF03207801
23. Intille, S., Haynes, C., Maniar, D., Ponnada, A., Manjourides, J.: μEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1124–1128. UbiComp 2016, Association for Computing Machinery (2016). https://doi.org/10.1145/2971648.2971717
24. Intille, S.S., Rondoni, J., Kukla, C., Ancona, I., Bao, L.: A context-aware experience sampling tool. In: CHI 2003 Extended Abstracts on Human Factors in Computing Systems. pp. 972–973. CHI EA 2003, Association for Computing Machinery (2003). https://doi.org/10.1145/765891.766101
25. Kaptein, M., Markopoulos, P., de Ruyter, B., Aarts, E.: Personalizing persuasive technologies: explicit and implicit personalization using persuasion profiles. Int. J. Hum Comput Stud. **77**, 38–51 (2015)
26. Khanshan, A.: Experiencer, the experience sampling method software (2021). https://experiencer.eu/
27. Khanshan, A.: Experiencer ESM source code (2022). https://github.com/khnshn/Experiencer
28. Khanshan, A.: From simulation to reality and back again: a hybrid approach to estimate the compliance of ESM study participants to different ESM protocols. In: 14th ACM SIGCHI Symposium on Engineering Interactive Computing Systems Doctoral Consortium, EICS DC 2022. 21–24 June 06 2022 (2022). http://eics.acm.org/eics2022/submission_dc.html
29. Khanshan, A., Van Gorp, P., Nuijten, R., Markopoulos, P.: Assessing the influence of physical activity upon the experience sampling response rate on wrist-worn devices. Int. J. Environ. Res. Public Health **18**(20), 10593 (2021). https://doi.org/10.3390/ijerph182010593
30. Kheirkhahan, M., et al.: A smartwatch-based framework for real-time and online assessment and mobility monitoring. J. Biomed. Inf. **89**, 29–40 (2019). https://doi.org/10.1016/j.jbi.2018.11.003
31. Kini, S.: Please take my survey: compliance with smartphone-based EMA/ESM studies (2013). https://digitalcommons.dartmouth.edu/senior_theses/83/
32. Larson, R., Csikszentmihalyi, M.: The experience sampling method. New Dir. Methodol. Soc. Behav. Sci. **15**, 41–56 (1983)
33. Larson, R., Csikszentmihalyi, M.: The experience sampling method. In: Flow and the Foundations of Positive Psychology, pp. 21–34. Springer, Dordrecht (2014). https://doi.org/10.1007/978-94-017-9088-8_2
34. Manini, T.M., et al.: Perception of older adults toward smartwatch technology for assessing pain and related patient-reported outcomes: pilot study. JMIR Mhealth Uhealth **7**(3), e10044 (2019). https://doi.org/10.2196/10044
35. Morren, M., van Dulmen, S., Ouwerkerk, J., Bensing, J.: Compliance with momentary pain measurement using electronic diaries: a systematic review. Eur. J. Pain **13**(4), 354–365 (2009). https://doi.org/10.1016/j.ejpain.2008.05.010

36. Myin-Germeys, I., Birchwood, M., Kwapil, T.: From environment to therapy in psychosis: a real-world momentary assessment approach. Schizophrenia Bull. **37**(2), 244–247 (2011). https://doi.org/10.1093/schbul/sbq164
37. Nuijten, R., et al.: Health promotion through monetary incentives: evaluating the impact of different reinforcement schedules on engagement levels with a mHealth app. Electronics **10**(23), 2935 (2021)
38. Park, S., Constantinides, M., Aiello, L.M., Quercia, D., Van Gent, P.: WellBeat: a framework for tracking daily well-being using smartwatches. IEEE Internet Comput. **24**(5), 10–17 (2020). https://doi.org/10.1109/MIC.2020.3017867
39. Pejovic, V., Lathia, N., Mascolo, C., Musolesi, M.: Mobile-based experience sampling for behaviour research. In: Tkalčič, M., De De Carolis, B., de de Gemmis, M., Odić, A., Košir, A. (eds.) Emotions and Personality in Personalized Services. HIS, pp. 141–161. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-31413-6_8
40. Ponnada, A., Thapa-Chhetry, B., Manjourides, J., Intille, S.: Measuring criterion validity of microinteraction ecological momentary assessment (micro-EMA): Exploratory pilot study with physical activity measurement. JMIR Mhealth Uhealth **9**(3), e23391 (2021). https://doi.org/10.2196/23391
41. Ram, N., Brinberg, M., Pincus, A.L., Conroy, D.E.: The questionable ecological validity of ecological momentary assessment: considerations for design and analysis. Res. Hum. Dev. **14**(3), 253–270 (2017). https://doi.org/10.1080/15427609.2017.1340052
42. Ranjan, Y., et al.: RADAR-base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. JMIR Mhealth Uhealth **7**(8), e11734 (2019). https://doi.org/10.2196/11734
43. Redmiles, E.M., Bodford, J., Blackwell, L.: I just want to feel safe: a diary study of safety perceptions on social media. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, pp. 405–416 (2019)
44. Rough, D.J., Quigley, A.: End-user development of experience sampling smartphone apps-recommendations and requirements. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **4**(2), 1–19 (2020)
45. Shahrestani, A., Van Gorp, P., Le Blanc, P., Greidanus, F., de Groot, K., Leermakers, J.: Unified health gamification can significantly improve well-being in corporate environments. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4507–4511 (2017). https://doi.org/10.1109/EMBC.2017.8037858. ISSN: 1558-4615
46. Shin, D.H., Biocca, F.: Health experience model of personal informatics: the case of a quantified self. Comput. Hum. Behav. **69**, 62–74 (2017). https://doi.org/10.1016/j.chb.2016.12.019
47. Singh, G., Delamare, W., Irani, P.: D-SWIME: a design space for smartwatch interaction techniques supporting mobility and encumbrance. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–13. Association for Computing Machinery (2018). https://doi.org/10.1145/3173574.3174208
48. Stone, A.A., Shiffman, S.: Ecological momentary assessment (EMA) in behavorial medicine. Ann. Behav. Med. **16**(3), 199–202 (1994). https://doi.org/10.1093/abm/16.3.199
49. Timmermann, J., Heuten, W., Boll, S.: Input methods for the borg-RPE-scale on smartwatches. In: 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), pp. 80–83 (2015). https://doi.org/10.4108/icst.pervasivehealth.2015.259220. ISSN: 2153-1641

50. Trull, T.J., Ebner-Priemer, U.W.: Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: Introduction to the special section. - PsycNET (2009). https://doi.org/10.1037/a0017653
51. Van Gorp, P.: Gamebus - social health games for the entire family (2015). http://www.gamebus.eu
52. Van Gorp, P., Surendranathan, A., Lesani, Y.: GameBus API guide (2021). http://devdocs.gamebus.eu
53. Wang, X., et al.: HOPES: an integrative digital phenotyping platform for data collection, monitoring, and machine learning. J. Med. Internet Res. **23**(3), e23984 (2021). https://doi.org/10.2196/23984
54. Wen, C.K.F., Schneider, S., Stone, A.A., Spruijt-Metz, D.: Compliance with mobile ecological momentary assessment protocols in children and adolescents: a systematic review and meta-analysis. J. Med. Internet Res. **19**(4), e6641 (2017). https://doi.org/10.2196/jmir.6641
55. Wheeler, L., Reis, H.T.: Self-recording of everyday life events: origins, types, and uses. J. Pers. **59**(3), 339–354 (1991). https://doi.org/10.1111/j.1467-6494.1991.tb00252.x
56. Zhang, C., Wang, S., Aarts, H., Dastani, M.: Using cognitive models to train warm start reinforcement learning agents for human-computer interactions. arXiv:2103.06160 (2021). arxiv.org/abs/2103.06160

BONVITA: Enabling Integrated Self-Care for Improving Chronic Patient's Wellbeing

Haridimos Kondylakis(✉), Angelina Kouroubali, and Dimitrios G. Katehakis

Institute of Computer Science, Foundation for Research and Technology - Hellas, N. Plastira 100, Vassilika Vouton, 700 13 Heraklion, Crete, Greece
`{kondylak,kouroub,katehaki}@ics.forth.gr`

Abstract. In 2020, more than 20% of the population in the European Union (EU) was over 65 years of age, according to the statistical office of the EU (eurostat), while the percentage of older people in cities in the industrialized part of the world is expected to pass this level in the following years. This demographic challenge stresses healthcare systems and requires novel self-management solutions. Several apps in use for the management of chronic diseases have the potential, if used in a systematic manner, to provide the necessary means, not only for early prediction and prevention of health deterioration, but also to support evidence based medicine through the sharing of data for secondary use. This paper presents the BONVITA solution, designed specifically for enabling integrated self-care for improving chronic patient's wellbeing through the provision of a series of visualization options, coherent consent management, and cohort formulation and analysis over available datasets. Despite the fact that it focuses on three chronic conditions (i.e., cardiovascular diseases (CVD), diabetes and chronic obstructive pulmonary disease (COPD)) its modular design enables its deployment into other chronic conditions as well, towards building a scalable and sustainable solution for both healthcare and social care that can be transferable to larger city/ region/ country contexts.

Keywords: Consent management · Integrated care · Patient Empowerment · Wellbeing

1 Introduction

Integrated care is an organizational principle for care delivery, aimed at achieving improved patient care, through better coordination of the provided services [1, 2]. According to Contandriapoulos et al. [3] integrated health services are health services managed and delivered so that people receive a continuum of services relevant to health promotion, disease prevention, diagnosis, treatment, disease-management, rehabilitation and palliative care. These services are coordinated across the different levels and sites of care within and beyond the health sector, and according to people's needs throughout their life course.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 332–343, 2023.
https://doi.org/10.1007/978-3-031-34586-9_22

As the global population ages, the risk for developing a number of chronic diseases such as diabetes, hypertension, arthritis or other heart and/ or respiratory disorders increases [4]. Supportive physical and social environments enable people to do what is important to them, despite capacity losses [5].

What is considered to be important for the patient is to be in a position to plan her/ his care with people who work together to understand her/ him and her/his carer(s) [6]. Allowing control and bringing together services to achieve outcomes that are important to the patient is imperative.

Integrated care should not be solely regarded as a means to managing medical problems since the principles extend to the wider definition of promoting health and wellbeing [7]. It is important to note that according to [8] in 2021 the share of those aged 65 to 79 was higher in rural regions (16%) in the European Union (EU) than in urban regions (14%), in contrast to working age population (aged 20–64), where there is a higher share in urban regions (60%) than in rural regions (58%).

E-health promises to provide comprehensive treatment of chronic patients in rural areas, by means of innovative tools (such as Electronic Medical Records (EMRs), patient portals, telehealth and personal health records (PHRs)) and state of the art technologies (such as Artificial Intelligence (AI), cybersecurity, and High Performance Computing (HPC)) to support innovative integrated models for self-care from home, powered by a large diversity of personal data, and supported by an ecosystem for wellbeing to help motivate patients and engage a large community of support around the patient. Although several infrastructures and mobile apps have been developed so far for enabling integrated self-care for improving chronic patient's wellbeing, they usually focus on a single domain such as mental health [9], stress [10], frailty [11], COVID-19 [12], cancer [13, 14] etc. However, an open infrastructure is required in order to facilitate the common needs among all chronic diseases, enabling also the fine-grained service delivery per chronic disease.

Key considerations in this direction, have to do with compliance with relevant data protection regulations (such as the EU General Data Protection Regulation (GDPR)) and easy to manage consent management for accessing personal health data, link to nontraditional health data (such as exercise goals and eating habits), and proactive preventive care. All of this requires secure and timely access and/or sharing of data between organizations. Healthcare resources should be proactively ready to be used (only) when a chronic patient is in need, while a supportive environment and data-driven digital tools should create a supportive environment capable of preserving health and the self-management of citizens.

This paper proposes a solution that includes services beneficial for both healthcare and social services to support rural areas moving beyond data silos of a single chronic disease and enabling citizens to open and trustworthy control their own data and the use of it. It focuses on improving self-management ability, as well as building a wider ecosystem for well-being including health and social care providers. It combines self-monitoring care in conjunction to the ecosystem of well-being towards the delivery of tailored advice, based not only on clinical findings, but also on genetic profiles, privately collected data, registry data, and other publicly available data relevant to the condition of the patient under consideration, in line with the key objectives of European Health Data Space (EHDS) [15]. The paper focuses on the design of this novel platform, encompassing current needs for integrated care for use, especially in rural areas, where the demographic challenge is higher, while at the same time is transferable and scalable to larger city contexts. To prove the scalability and the extensibility of the platform it will support disease-specific apps for cardiovascular diseases (CVD), Chronic obstructive pulmonary disease (COPD) and diabetes. The platform relies on already available modules implemented for stress (through the STARS-PCP project [9]) and frailty management (through the eCare PCP project [10]) repurposed for enabling integrated self-care for improving chronic patient's wellbeing.

The rest of this paper is structured as follows: In Sect. 2 we focus on use cases available and requirement analysis and Sect. 3 we present the designed architecture and the available modules. Finally, Sect. 4 concludes the paper.

2 Requirement Analysis and Use Cases

To enhance autonomy and support independent living, it is important to encourage citizens to take responsibility of their own health needs [16]. This is facilitated through technologies that provide digital services for accessing personal health data, gaining insights about health conditions and controlling access to data through notions of GDPR legislation and the new European data strategy [17, 18]. Health data ownership facilitates citizens to control their data and allow access to various actors based on their health needs but also as part of a volunteer offering to their data to the research communities [19]. Access to health data allows for a range of new possibilities to create systems that offer better control of health data and the ability to share data when appropriate. Data sharing can facilitate research towards personalized care approaches as well as for the creating of effective interventions to improve health and well-being [2].

Organizational structures for storing health data at healthcare and social care institutions are not focusing on how to facilitate the utilization of these data. Governed by legislation, health data remains in silos and cannot be shared for integrated care or to inform and guide research [20]. Moreover, health and social care systems have different structures and are rarely integrated. As a result, the actual technologies for self-monitoring do not allow for the sharing of data amongst healthcare and social service providers. New model designs that include sharing of data that has been generated by different actors, are imperative. Therefore, it is necessary to involve critical actors including citizens, eHealth and the welfare tech industry as well as healthcare and social service providers (private and public) [21].

Collected data needs to be collected and reused using cutting-edge innovation and technologies. A data lake can in a simple form serve as a safe harbour to store personal health data for citizens. An advance data lake can provide further benefits to different actors. It is applicable to different services as well as to different research, innovation and development activities [22, 23]. Transparent and open policies promote and build citizen trust. Feedback systems demonstrate gains for the individuals, both in the short and long run, and technological platforms that use the latest encryption technologies facilitate the development of a secure public-private data lake environment of data sharing [24].

The multidimensional needs of the individual would need more than just technology at home. Integrated care combined with self-management requires the interaction of health and social care services, in a flexible way to support individual needs and social context [25]. BONVITA as a technology platform will be tailored to improve control, safety, security, freedom and awareness of citizens' well-being.

Self-management of chronic diseases require citizens to have access to technologies that facilitate data management, provide tools to enhance health literacy about their condition as well as provide motivation based on their personal preferences [26]. Such technologies empower citizens to take responsibility for the design of their individual integrated care model and become self-responsible for their own health. This empowerment will improve patient's self-esteem making the system fully succeed [27].

Technology platforms need to be based on a value-based model revolving around the patient's needs to meet its main aim of improving patients' wellbeing [16]. For this reason, a health technology assessment (HTA) framework will be applied based on the early engagement of stakeholders, systematic literature reviews and scenario analysis to estimate the potential value of BONVITA, the digital platform which is under development [28, 29]. To establish the value and the impact of the technology, all involved stakeholders will be engaged during the evaluation phase including health and social care professionals, patients, informal carers, and administrative personel [30]. The framework evaluates stakeholder values on four domains (patient, economic, clinical and organisational) at defined stages with the purpose of:

- Identifying unmet user needs, elements adding value to the system and those to be modified.
- Generating evidence on results
- Evaluating the performance of suppliers and solutions.
- Identifying areas of opportunities
- Helping the innovators to complete a product that the health service wants

However, analysis of the details of the HTA framework that will be applied is beyond the scope of this paper. The following section focuses on the technical architecture of the proposed solution.

3 Architecture

The architecture of the proposed solution is shown in Fig. 1 and consists of four layers. The applications tier with mobile, disease-specific apps, the business logic tier with the various services provided the data management and finally the computational tier. In the sequel we will analyze each of those layers in detail.

3.1 Application Tier

The BONVITA infrastructure aims at enabling citizens to control their own health data, integrating data driven services for the self-management of three chronic diseases, i.e.,

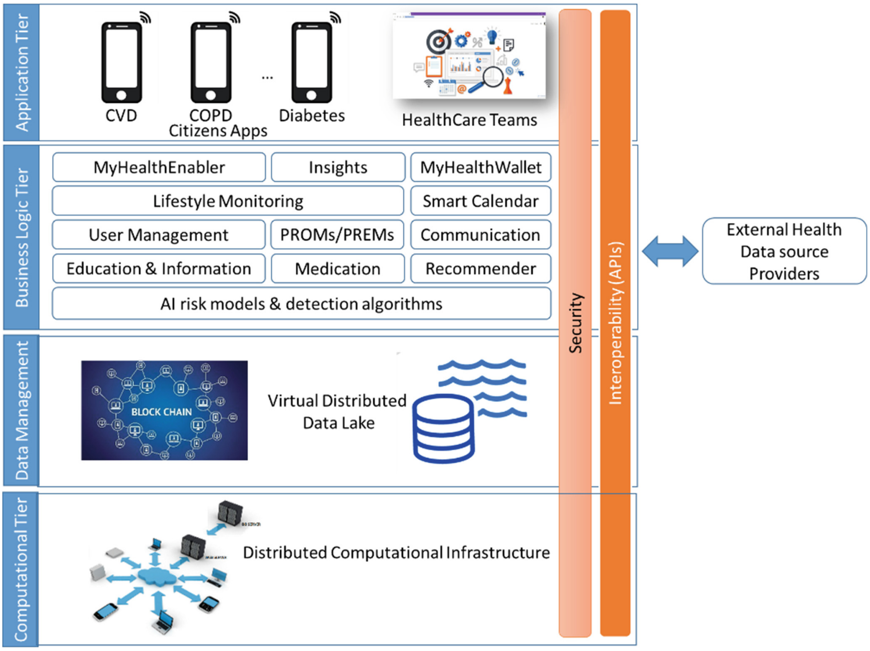


Fig. 1. The high-level architecture for the BONVITA solution for an integrated self-management model to improve chronic patient's wellbeing

COPD, Diabetes and CVD, in rural areas, selected due to their high prevalence and the fact that significant evidence exists for the positive results in self-management:

- **CVD:** A monitoring system for citizens at risk can prevent and/or predict adverse events, enabling better coordinate care. Further a CVD app can help citizens in adopting a healthier lifestyle and help them to adapt their changes to real outcomes.
- **Diabetes:** Diabetes is one of the most complex chronic conditions for patients to deal with, especially when it is in an advanced state or it is type 1, as it requires significant changes in the lifestyle.
- **COPD:** Mobile apps focusing on this disease can help educate citizens on how to live with their disease and how to rehabilitate themselves by reducing the visits to the hospitals.

Although those three diseases have been selected for proof-of-concept deployment of the BONVITA services, other disease-specific apps are also supported as well by the BONVITA ecosystem that helps motivate patients and engage them through a large community of support, enabling self-monitoring and self-care from home. All apps are able to capture citizen reported outcomes over time using secure, anonymous but identifiable data.

3.2 Business Logic Tier

Next, we present the key modules of the business logic tier. Those are the following:

MyHealthEnabler module provides a series of visualization options for individual health data. In essence it accesses the available health information and provides a pallete of graphical diagrams that can visualize the information at granular level, enabling zooming operations as well.

MyHealthWallet module provides consent management and sharing of the data between users of the various applications. Citizens are the owners of their data and can select to share their data with other users or healthcare providers. That consent can be easily withdrawn at will, whereas access to the information is recorded.

Insight's module provides cohort formulation and analysis over the available datasets enabling analysts to gain useful insights on the store data. An overview of the data sets is provided, together with a data dictionary wherever applicable referring to relevant standard semantic definitions, clarifying intended meaning of data, including used algorithms in the case of computed data. Through the insights module data can be selected, anonymized, and extracted for research purposes.

Lifestyle Monitoring module collects user information about vital signs and biometric data, physical activity, food and nutrition, psychological state of mind, therapy adherence, and socialization. The implemented APIs enable the ingestion of various types of vital signs and can be pushed by mobile applications through their connected connected devices such as digital blood pressure monitors, digital glucose monitors, digital pulse oximeters, sleep and stress monitors, etc. In addition, the lifestyle monitoring module can collect user biometric data retrieved through smartwatch sensors (heart rate, sleep quality, and physical activity). Further vital sign can be recorded manually by the various mobile apps through the appropriate interfaces they implement as still the same API can be used for data recording. All vital signs available in the data layer can be interactively visualized through the MyHealthEnabler and the Insights modules for both the patients and the clinicians. A screenshot of the clinician view is shown in Fig. 2.

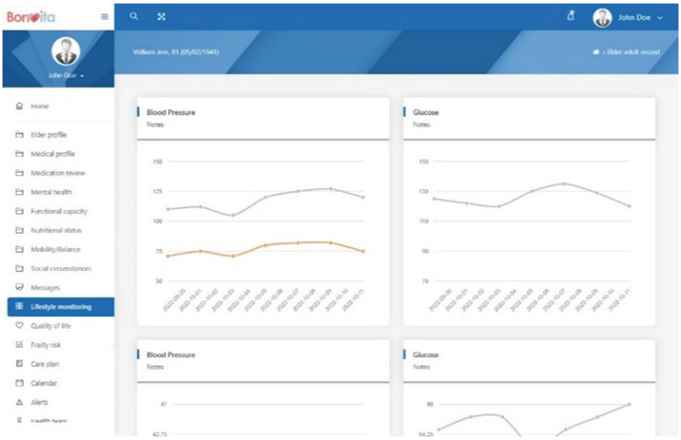


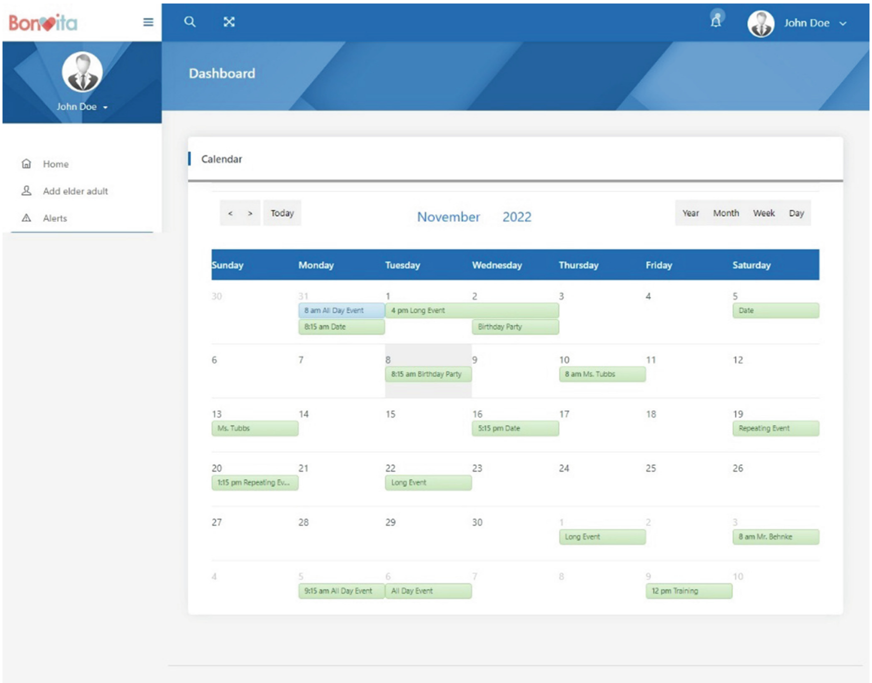
Fig. 2. Visualization of the vital signs stored in the data layer

Medication's module enables the management of the various medications that the citizens might use, enabling polypharmacy management, drug interaction detection and alert etc.

PROMs/PREMs are implemented **through the self-assessment** module. Patients use this module to report proms and prems in collaboration with their healthcare providers in order to collect relevant information from the citizens.

User Management module differentiates user types and can include new users in the system, including patients, their caregivers, formal caregivers, healthcare professionals, and social workers. An administration panel enables the creation of new healthcare professional and/or social worker profiles.

The Smart Calendar module enables patients, the caregivers, and care professionals to set up different events in the foreseeable future (e.g., planned goals and activities before the next meeting, which itself is planned and entered into the schedule). The patient can define personal reminders such as family birthdays and other socialization-related events as well. A screenshot of the smart calendar is shown in Fig. 3 and is connected with the various categories of information available for both the patient and the healthcare providers.

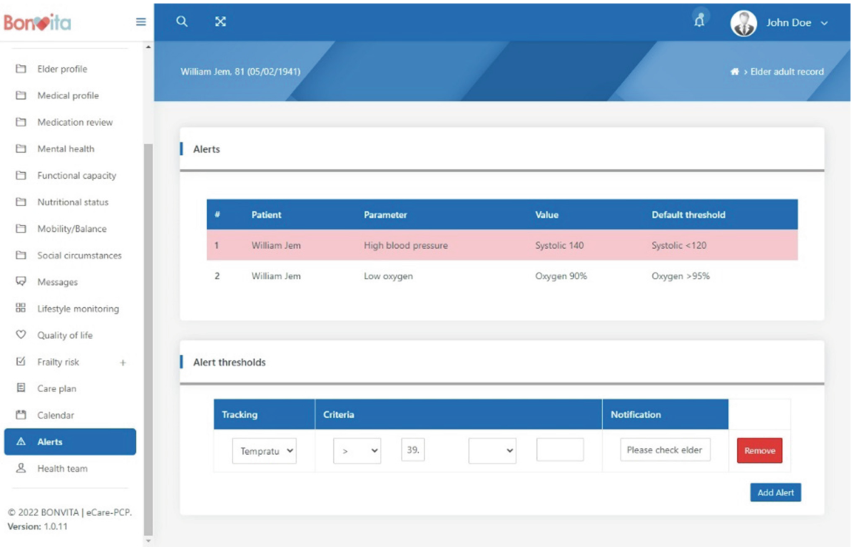


Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
30	31 8 am All Day Event 8:15 am Date	1 4 pm Long Event	2 Birthday Party	3	4	5 Date
6	7	8 8:15 am Birthday Party	9	10 8 am Ms. Tubbs	11	12
13 Ms. Tubbs	14	15	16 5:15 pm Date	17	18	19 Repeating Event
20 1:15 pm Repeating Ev...	21	22 Long Event	23	24	25	26
27	28	29	30	1 Long Event	2	3 8 am Mr. Behnke
4	5 9:15 am All Day Event	6 All Day Event	7	8	9 12 pm Training	10

Fig. 3. Smart Calendar visualization

Communication module supports synchronous and asynchronous communication between patients and their care team, as well as between citizens for social purposes, by exploiting a messaging system. Predefined messages can be configured to be sent to users at specific time points based on the shared care plan. Further the communication module facilitates access to existing peer networks by grouping accounts in forum like places, enabling to exchange opinions, and experiences.

Recommender module sends personalized pieces of advice about how to reach the level of health behaviors that each user requires. This module is directly linked with the vital signs and activity monitoring module, whereas it is possible to recommend also activities based on user profiles (e.g., guided walks). Currently using the recommender module the healthcare professionals can set recommendation rules based on the available data, specifying which actions should be performed with given conditions [31]. The module is also used for specifying alerts that should notify the patients or the healthcare providers with a critical flag as well. A screenshot of the alerts shown in the healthcare provider view is shown in Fig. 4.



Alerts

#	Patient	Parameter	Value	Default threshold
1	William Jem	High blood pressure	Systolic 140	Systolic <120
2	William Jem	Low oxygen	Oxygen 90%	Oxygen >95%

Alert thresholds

Tracking	Criteria				Notification
Tempratu	>	39.			Please check elder

Fig. 4. Alerts from the clinician view

Education & Information module provides information material, educational resources, and training to support emotional, psychological, and physical well-being and empowerment. Training materials are provided in order to increase the skills of citizens as well as their caregivers and care professionals. The training content is selected from a list of available modules divided in three main topics: (i) soft skills for training emotion regulation strategies, effective communication and active listening, empathy, negotiation, and conflict resolution, resilience and self-motivation, loss and grief, etc.; (ii) IT usage training on new technologies in order to leverage the potential of the system itself as well as to empower patients and care professionals; and (iii) training in acts of care such as self-care, health promotion, care techniques, adoption of healthier lifestyles etc. The training modules will include multi-media content (tutorials, videos, interviews, etc.) with special attention on demonstrations (e.g., how to use medical devices). The presented content is dynamic, i.e., content is suggested depending on the user's knowledge and capabilities as well as their accomplishments during training.

Decision Support System module will implement AI risk models and detection algorithms for calculating the various clinically validated risk scores available in the BONVITA platform based on clinically validated models (e.g., Seattle risk model for CVD). Besides these risk scores, machine learning algorithms will continuously monitor all data that are collected trying to identify dominant features and the most influential factors for risks, including also the various scores calculated and the healthcare professionals' evaluations. We foresee that as the multidimensional data collected through the BONVITA solution grow, we will be able to build models with better accuracy with respect to existing risk scores, tailored specifically to the citizens using each health app.

3.3 Data Management Tier

All data available within the BONVITA solution will be stored centrally in a virtual data lake, exploiting the OMOP-CDM and the FHIR data model for storage. In addition, relevant terminologies such as SNOMED-CT, LOINC, MEDRA, ICD-10 etc. will enable the unambiguous storage of the various data. As such the data will be already FAIRified, I.e., findable, interoperable, reusable and accessible through APIs that will provide the necessary data to the various modules. The virtual data lake will allow applications to blend data from various sources, local or remote, and it is connected or disconnected from data sources when required by the apps that use it. It is a directory of citizens' data sources (with accurate and updated data descriptions) so that applications can search for data as if it were a single data set and retrieve the needed data from its location. It guarantees control access based on user consent (managed by Smart Contracts). The users of applications/analysts/insights may only see and/or access the data they are entitled to.

3.4 Computational Power Tier

The bottom layer is the actual computing nodes running the different components of the entire system. The computational layer is distributed enabling scalability and elasticity upon request, whereas each node is intended to be operated by independent organizations, and each node consists of both a public and a private layer.

3.5 Security and Interoperability

Security spans across all layers whereas interoperability services are offered by both the data management & AI tier and the business logic tier as well.

Regarding security, **blockchain is** used for storing all relevant information in the data lake. Smart contracting is used for making changes in the data enabling effective logging of. Further **encryption** is used for pseudonymization, data integrity, confidentiality and accountability.

In ensuring successful **interoperability** the BONVITA solution, key technical syntactic & semantic standards are supported. For the healthcare domain, ICD10–11

CM, LOINC, SNOMED CT, DICOM, HL7/FHIR and IEEE-073 PHD/ITU H.813 are adopted. For the data exchange of these health-related data sets, it should be using HL7/FHIR is adopted for data exchange of health-related datasets between the internal and external systems.

4 Discussion and First Results

One of the main novelties of the designed infrastructure is that it breaks the silos of disease specific mobile apps and enables the harmonic management of all health information of the citizens. Further, it allows the sharing of data through smart contracts and consent, enabling healthcare providers to have regulated access to the available information. The infrastructure has already been validated in scenarios for the management of stress [10], where healthcare professionals evaluated BONVITA in terms of functionality, design, clearness of the instructions of use, and general usability. In detail, about functionality and design, healthcare professionals positively evaluated BONVITA as useful, pleasant, easy to use and clear. Similar positive evaluations have also been reported on the clearness of the instructions of use and BONVITA solution has been evaluated as recommend to other patients and healthcare professionals. Nevertheless, these positive results, mixed evaluations have been reported regarding the BONVITA ability to communicate with patients and as added value. Currently the BONVITA solution is being evaluated for the management of frailty and through the following months it will be tested first by healthcare providers and then by elder adults.

5 Conclusions

This paper presents the design of a novel platform for enabling integrated self-care for improving chronic patients' wellbeing. The platform capitalizes modules already implemented for stress and frailty management, repurposed through a modular infrastructure able to accommodate management for multiple chronic diseases. We anticipate that in the following months the platform will be linked with other external apps for managing COPD, CVD and diabetes exploiting the rich infrastructure already available. Key challenges that we expect to be successfully addressed by its use, under different settings, concern the degree of openness of the platform for the incorporation of personal data spaces, the effectiveness of leveraging access to real-world data, the effective coordination of caregivers, and the empowerment of citizens in a trustworthy environment.

References

1. What is integrated care? https://www.nuffieldtrust.org.uk/files/2017-01/what-is-integrated-care-report-web-final_copy1.pdf. Accessed 10 Apr 2022
2. Katehakis, D.G., et al.: Integrated care solutions for the citizen: personal health record functional models to support interoperability. Ejbi **13**(1), 41–56 (2017)
3. Contandriapoulos, A.P., Denis, J.L., Touati, N., Rodriguez, C.: The integration of health care: Dimensions and implementation, Montréal: Université de Montréal; Jun, Groupe de recherche interdisciplinaire en santé. Working Paper N04–01 (2003)

4. Implications of Aging. https://muschealth.org/medical-services/geriatrics-and-aging/healthy-aging/implications-of-aging. Accessed 04 Oct 2022
5. Aging and Health. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health. Accessed 04 Oct 2022
6. Lewis, R., Rosen, R., Goodwin, N., Dixon, J.: Where next for integrated care organisations in the English NHS? The Nuffield Trust, London (2010)
7. Goodwin, N.: Understanding integrated care. Int. J. Integr. Care, **6**, 4 2016
8. Rural and urban regions: differences. https://ec.europa.eu/eurostat/cache/digpub/demography/bloc-3d.html?lang=en. Accessed 04 Oct 2022
9. Kieran, W., et al.: Beyond mobile apps: a survey of technologies for mental well-being. IEEE Trans. Affect. Comput. **13**, 1216–1235 (2020)
10. Kondylakis, H., et al.: A digital health intervention for stress and anxiety relief in perioperative care: protocol for a feasibility randomized controlled trial. JMIR Res. Protoc. **11**, 38536 (2022)
11. Sykoutris, A., et al.: iCompanion: a serious games app for the management of frailty. In: Challenges of Trustable AI and Added-Value on Health, pp. 624–628 (2022)
12. Kondylakis, H., et al.: COVID-19 mobile apps: a systematic review of the literature. J. Med. Internet Res. **22**(12), e23170 (2020)
13. Kondylakis, H., et al.: Status and recommendations of technological and data-driven innovations in cancer care: focus group study. J. Med. Internet Res. **22**, e22034 (2020)
14. Kouroubali, A., et al.: An integrated approach towards developing quality mobile health apps for cancer. In: Mobile Health Applications for Quality Healthcare Delivery. IGI Global, pp. 46–71 (2019)
15. Proposal for a regulation - The European Health Data Space. https://health.ec.europa.eu/publications/proposal-regulation-european-health-data-space_en. Accessed 04 Oct 2022
16. Kouroubali, A., Chiarugi, F.: Developing advanced technology services for diabetes management: user preferences in Europe. In: Nikita, K.S., Lin, J.C., Fotiadis, D.I., Arredondo Waldmeyer, M.-T. (eds.) MobiHealth 2011. LNICSSITE, vol. 83, pp. 69–74. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29734-2_10
17. Kouroubali, A., Katehakis, D. G.: Policy and strategy for interoperability of digital health in Europe. In: MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation, pp. 897–901. IOS Press (2022)
18. Kouroubali, A., Katehakis, D.G.: The new European interoperability framework as a facilitator of digital transformation for citizen empowerment. J. Biomed. Inform. **94**, 103166 (2019)
19. Kouroubali, A., Kondylakis, H., Katehakis, D.G.: Integrated care in the Era of COVID-19: turning vision into reality with digital health. Frontiers Digital Health, **3**, 647938 (2021)
20. Kondylakis, H., et al.: Developing a data infrastructure for enabling breast cancer women to BOUNCE back. In: CBMS, pp. 652–657. IEEE (2019)
21. Kondylakis, H., et al.: Status and recommendations of technological and data-driven innovations in cancer care: focus group study. J. Med. Internet Res. **22**(12), e22034 (2020)
22. Kondylakis, H., et al.: CareKeeper: a platform for intelligent care coordination. In: BIBE, pp. 1–4 (2021)
23. Ravat, F., Zhao, Y.: Data lakes: trends and perspectives. In: Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) DEXA 2019. LNCS, vol. 11706, pp. 304–313. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27615-7_23
24. Stern, A.D., et al.: Advancing digital health applications: priorities for innovation in real-world evidence generation. Lancet Digital Health **4**(3), e200–e206 (2022)
25. Kondylakis, H., Kouroubali, A., Koumakis, L., Rivero-Rodriguez, A., Hors-Fraile, S., Katehakis, D.G.: Designing a novel technical infrastructure for chronic pain self-management. In: PHealth, vol. 203, p. 249. IOS Press (2018)

26. Kouroubali, A., Kondylakis, H., Koumakis, L., Papagiannakis, G., Zikas, P., Katehakis, D.G.: iSupport: building a resilience support tool for improving the health condition of the patient during the care path. pHealth, 261, 253–258 (2019)
27. Yardley, L., Morrison, L., Bradbury, K., Muller, I.: The person-based approach to intervention development: application to digital health-related behavior change interventions. J. Med. Internet Res. **17**(1), e4055 (2015)
28. Mateo-Abad, M., et al.: Impact assessment of an innovative integrated care model for older complex patients with multimorbidity: the CareWell project. Int. J. Integr. Care, 20(2) (2020)
29. Larrañaga, I., Stafylas, P., Fullaondo, A., Apuzzo, G.M., Mar, J.: Economic evaluation of an integrated health and social care program for heart failure through 2 different modeling techniques. Health Serv. Res. Manag. Epidemiol. **5**, 2333392818795795 (2018)
30. Wale, J.L., Thomas, S., Hamerlijnck, D., Hollander, R.: Patients and public are important stakeholders in health technology assessment but the level of involvement is low–a call to action. Res. Involvement Engagem. **7**(1), 1–11 (2021)
31. Kondylakis, H., et al.: Semantically-enabled personal medical information recommender. In: ISWC (Posters & Demos) (2015)
32. STARS-PCP EU project. https://stars-pcp.eu/. Accessed 04 Oct 2022
33. eCARE-PCP EU project. https://ecare-pcp.eu/. Accessed 04 Oct 2022

Data Analytics for Health and Connected Care: Ontology, Knowledge Graph and Applications

Bram Steenwinckel[1](✉), Mathias De Brouwer[1], Marija Stojchevska[1],
Jeroen Van Der Donckt[1], Jelle Nelis[1], Joeri Ruyssinck[2],
Joachim van der Herten[2], Koen Casier[3], Jan Van Ooteghem[3],
Pieter Crombez[4], Filip De Turck[1], Sofie Van Hoecke[1], and Femke Ongenae[1]

[1] IDLab, Ghent University-imec, Technologiepark 126, 9050 Gent, Belgium
`bram.steenwinckel@ugent.be`
[2] ML2Grow, Reigerstraat 8, 9000 Gent, Belgium
`info@ml2grow.com`
[3] Amaron, Kapellestraat 13, 8755 Ruiselede, Belgium
`info@amaron.be`
[4] Televic Healthcare, Leo Bekaertlaan 1, 8870 Izegem, Belgium
`healthcare@televic.com`

Abstract. Connected care applications are increasingly used to achieve a more continuous and pervasive healthcare follow-up of chronic diseases. Within these applications, objective insights are collected by using Artificial Intelligence (AI) models on Internet of Things (IoT) devices in patient's homes and by using wearable devices to capture biomedical parameters. However, to enable easy re-use of AI applications trained and designed on top of sensor data, it is important to uniformly describe the collected data and how this links to the health condition of the patient. In this paper, we propose the DAHCC (Data Analytics For Health and Connected Care) ontology, dataset and Knowledge Graph (KG). The ontology allows capturing the metadata about the sensors, the different designed AI algorithms and the health insights and their correlation to the medical condition of the patients. To showcase the use of the ontology, a large dataset of 42 participants performing daily life activities in a smart home was collected and annotated with the DAHCC ontology into a KG. Three applications using this KG are provided as inspiration on how other connected care applications can utilize DAHCC. The ontology, KG and the applications are made publicly available at https://dahcc.idlab.ugent.be. DAHCC's goal is to integrate care systems such that their outcomes can be visualised, interpreted and acted upon without increasing the burden of healthcare professionals who rely on such systems.

Keywords: Connected Care · Open health data · Ontology

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 344–360, 2023.
https://doi.org/10.1007/978-3-031-34586-9_23

1 Introduction

From the perspective of healthcare professionals, our current healthcare system brings many challenges as well as opportunities for digital healthcare [4]. Each healthcare professional already takes care of multiple people as more and more in-person visits have to be reduced to an absolute minimum. Connected care solutions, including remote patient monitoring and secure communications between clinicians or caregivers and their patients, may rapidly become the first choice to provide care in a public health system [21].

First steps have already been taken to provide such connected care solutions. Smart cities are designed using smart sensors to track the well-being of their citizens [10], ambient intelligent homes where sensors track and derive information from their residents in a privacy-friendly manner are being deployed [5] and patients with specific healthcare-related problems are equipped with smart sensors, such as wearables, to track biometric properties [7].

Those care applications deliver insightful information for caregivers and health professionals and should eventually help them to work more effectively [25]. Artificial Intelligence (AI) applications, such as Human Activity Recognition (HAR), can be used to monitor the physical abilities of elderly [9]. The HAR outcome or prediction can be seen as new knowledge, which can be of interest to a healthcare professional [1].

However, monitoring and analysing all the data produced by intelligent sensors and care applications should not increase the burden and stress of those who rely on them [3]. Therefore, a uniform method is needed to both describe the connected care applications and associated data. By doing this, the derived knowledge from those care applications can be easily integrated into new smart applications and can be automatically reported using insightful dashboards. They can also be used in autonomous alerting tools to inform health care professionals when unwanted behaviour or situations of high risk occur.

In this paper, we present the Data Analytics for Health and Connected Care (DAHCC) Ontology, a method to semantically describe sensor data, human activities, smart applications, health actors, faulty or unwanted behaviour and link them all together. To facilitate the idea behind DAHCC and the creation of new connected care applications, an ambient intelligent dataset was created and semantically enriched into a Knowledge Graph (KG) with over 40 participants performing daily and nightly activities. Example applications show how new connected care applications can be developed based on this dataset. Both the used method (ontology), datasets, KG, as well as the used applications are made available online under an MIT license.

The remainder of this paper defines in Sect. 2 the different existing techniques to describe healthcare resources and applications uniformly. The creation of the DAHCC ontology based on the multiple reused industry standards and well-adopted ontologies is provided in Sect. 3. Section 4 describes the creation of the dataset and Sect. 5 details how it was transformed into a KG. The applications built upon this KG are described in Sect. 6.

2 Related Work

Monitoring of patients in either an ambient life setting or for general healthcare purposes covers a large research domain. The usage of IoT devices and wearables was crucial in the development of these applications [13]. In the last decade, healthcare ontologies became popular to incorporate and combine domain knowledge with these sensory devices. Table 1 summarises the field of connected care ontologies. Most of these ontologies are not publicly available to be reused in other applications or were designed for a specific use case within a connected care setting.

Table 1. Overview of existing Healthcare-related ontologies

Ontology	Available	Reused Ontologies	Semantically describe
ACCIO [17]	Yes	SSN	Ambient patient rooms
Telehealth [12]	No	ICD, ICF, SNOMED-CT	Patient profile
HealthIoT [20]	No	NaN	Medical devices
SAREF4EHAW [14]	Yes	SAREF	Monitoring health actors
e-Health [11]	No	SSN	Device communication
IFO [19]	No	NaN	IoT health and fitness data
LHR [18]	No	SNOMED-CT, SSN	Health care data exchange
SHCO [24]	Yes	SAREF	Patient-doctor interactions
Do-Care [8]	No	FOAF, SOSA, ICNP	Nurse interactions
DAHCC	**Yes**	**SAREF, EEP, OWL Time**	**Connected care**

The ACCIO ontology [17] exploits and integrates the heterogeneous data by utilising a continuous care ontology for patient rooms of the future in a hospital setting. It was one of the first ontologies co-created with nurses, caregivers, patients, doctors and professionals working in the healthcare industry. The ontology is used to exploit and integrate heterogeneous data. The Patient-Centric for Telehealth ontology [12] was designed with an explicit focus on the personality traits of the patients to describe diseases, functioning and physiological measurement. It does, however, not integrate sensor-related information. The HealthIoT ontology [20] aims to represent the semantic interoperability of the medical connected objects and their data. The ontology focuses only on the medical aspects of the devices and provides rules to analyse the detected vital signs. Those analyses can be delivered to a healthcare professional. SAREF4Health ontology [14] is an extension of SAREF, a framework for smart appliances references. The SAREF ontology describes how sensors and sensor data can be described and linked to each other. The SAREF4Health extension specifies eHealth/Ageing-well (EHAW) domain-related resources and defines how the sensor data relates to health actors. Care-specific needs or how new systems can provide additional information towards a connected care system are not provided in this

SAREF4Health extension. The e-Health ontology [11] tries to reduce programmatically implementing the interpretation of the data sender and data receiver for each new healthcare device added into a system. This ontology lowers the efforts needed to extend a current healthcare setup in an ambient living context. The IoT fitness ontology (IFO) [19] presents a semantic data model useful to consistently represent health and fitness data from heterogeneous IoT sources and integrate them into semantic platforms to enable automatic reasoning by inference engines. This ontology does not take into account standards such as SSN or SAREF to describe the sensor and data. Linked Health Resource (LHR) ontology [18] integrates health data from different services as linked resources. The healthcare-IoT ontology [22] provides semantic interoperability among heterogeneous devices and users in the healthcare domain. The ontology models the exchange of health care data and home environment data. The smart healthcare ontology (SHCO) [24] define concepts for monitoring doctors and patients anytime, anywhere. SHCO is presented as a semantic model by extracting healthcare knowledge such as doctor-patient records, recommended diagnoses and treatment policies. This ontology only uses rather static non-sensory information. The Do-Care ontology [8] is modular and incorporates the International Classification for Nursing Practice (ICNP) ontology together with inference rules. The methodology is dynamic and adjustable to meet possible changes in the medication market, medical discoveries, and personal users' profiles. Nurse interactions are the main focus of this ontology.

Each of these ontologies describes a clear subpart within the connected care domain. Almost all ontologies already reused existing ontologies such as SSN or SAREF. Some of them are also made open-source and are available online for further reuse. However, an ontology which describes all concepts related to connected care, ambient living, patient care and their interaction with healthcare professionals is currently not available. Moreover, there doesn't exist to our knowledge an ontology which links AI models, with their predictions and model configuration, to the monitored context or situation of a healthcare actor. Such an ontology is of high need as more and more AI models are being used within healthcare. Those models generate new insights that, when semantically described, can deliver even more advanced input to healthcare professionals. In this paper, the DAHCC ontology is created to resolve this need.

3 DAHCC Ontology Creation

The DAHCC ontology describes real-world connected care entities (person, sensor, activities, etc.) and their interrelationships. It also models domain knowledge, e.g., performing human activities & profile data or location information where those sensors are installed. The design of DAHCC ontology was based on three types of information sources:

- Structured documents, e.g. the descriptions of the used sensors and monitoring systems in technical specifications & APIs, in-take documents used by caregivers, etc.

- An assessment of existing ontologies that could be reused.
- Some knowledge is only available from the stakeholders, who perform day-to-day assessment & handling of deviating situations. To derive this information, decision tree workshops were organised with nurses and caregivers as described in [16].

The structured documents contained mostly resident-specific, privacy-sensitive information. We kept those fields that are interesting for healthcare-related cases. SAREF Core[1], SAREF4BLDG[2], SAREF4EHAW[3], SAREF4WEAR[4], Execution-Executor-Procedure (EEP)[5] and the OWL Time ontology[6] were reused within DAHCC. The decision tree workshops resulted in a long list of care intervention reasons, and an indication of which data/information the caregivers used to assess alarming situations and how they handle them. The information inside these decision trees was used to define new concepts regarding patient monitoring within the DAHCC ontology.

3.1 DAHCC Ontology

The DAHCC ontology exists out of five sub ontologies which have links to each other. Each of these ontologies reuses one or more of the mentioned, already existing ontologies, combined with new concepts derived from the provided workshops [16]. These workshops brought together ontologists, healthcare providers and people who monitor patient calls. The results of these workshops are identified use cases by the participants, a long list of call reasons and resulting decision trees of how care is provided or how a person monitors patient calls. These decision trees were consolidated by the ontologist into two decision trees: one for the nurses and one for the caregivers. The information inside these decision trees was used to define new concepts regarding patient monitoring and handling of call operations within the DAHCC ontology. For the extensive documentation of the ontological concepts, we refer the interested reader to the ontology section at https://dahcc.idlab.ugent.be. The remainder of this section describes the 5 different sub ontologies in the DAHCC dataset.

Activity Recognition. The Activity Recognition sub ontology defines how the concepts of activities, performed by a saref:HealthActor, can be predicted using an activity recognition model. The ontology also describes how such a model can be defined, together with its configuration and input and output data. The ontology describes more in general lifestyles, routines and anomaly classes for, e.g., unwanted daily or nightly activities. The Activity Recognition

---

[1] https://saref.etsi.org/core/v3.1.1/.
[2] https://saref.etsi.org/saref4bldg/v1.1.2/.
[3] https://saref.etsi.org/saref4ehaw/v1.1.1/.
[4] https://saref.etsi.org/saref4wear/v1.1.1/.
[5] https://iesnaola.github.io/EEP/index-en.html.
[6] https://www.w3.org/TR/owl-time/.

ontology will be used to describe all the concepts related to both performed and predicted activities, together with their link to the models/techniques used to detect them, the performed routines and the lifestyle of the health actor. An overview of this ontology is provided in Fig. 1.

Monitored Person. The Monitored Person sub ontology, as shown in Fig. 2, describes the person itself who is monitored by all sensors and who performs the human activities and routines. It also describes the possible diseases, addictions, mental illnesses or allergies this monitored person can have. Medication and the current mental state of the monitored person are also defined as concepts within this ontology. This Monitored Person ontology has a strong link to the Activity Recognition ontology to link human activities.

Sensors and Actuators. This Sensors and Actuators sub ontology describes a numerous amount of sensors and actuators that produce data and can be equipped inside a household. Besides the sensors and the measurement properties, this ontology also defines where those sensors are placed and which appliance or rooms they analyse. This subontology is shown in Fig. 3 and mainly extends the SAREF Core and SAREF4BLD ontologies with ambient life-specific concepts.

Sensors and Wearables. Similar to the Sensors and Actuators ontology, the Sensors and Wearables ontology (Fig. 4) describes the wearables and sensors that can be attached to or near a monitored person. This sub ontology extends the SAREF4WEAR ontology and adds specific connected care concepts to it. Wearables range from medical devices using simple near-field communication or Bluetooth connections to send medical parameters to a cloud environment, as well to more sophisticated smartwatches to track residents inside a building 24/7.

Caregiver. The Caregiver sub ontology defines the link between caregivers and monitored persons. It uses the SAREF4EHAW concepts to assume the possible relation between health actors. Additionally, it also describes how the required care can propagate to the responsible entity or caregiver. It also defines the Care Provisioning activity. An overview of this ontology is provided in Fig. 5.

4 DAHCC Dataset

To show how the DAHCC Ontology can be used to annotate healthcare data and how this is beneficial for the creation of connected care applications and extracting knowledge, a large ambient intelligent data collection campaign at the HomeLab of the IDLab research group of Ghent University was performed. The HomeLab is an actual standalone house offering a unique residential test

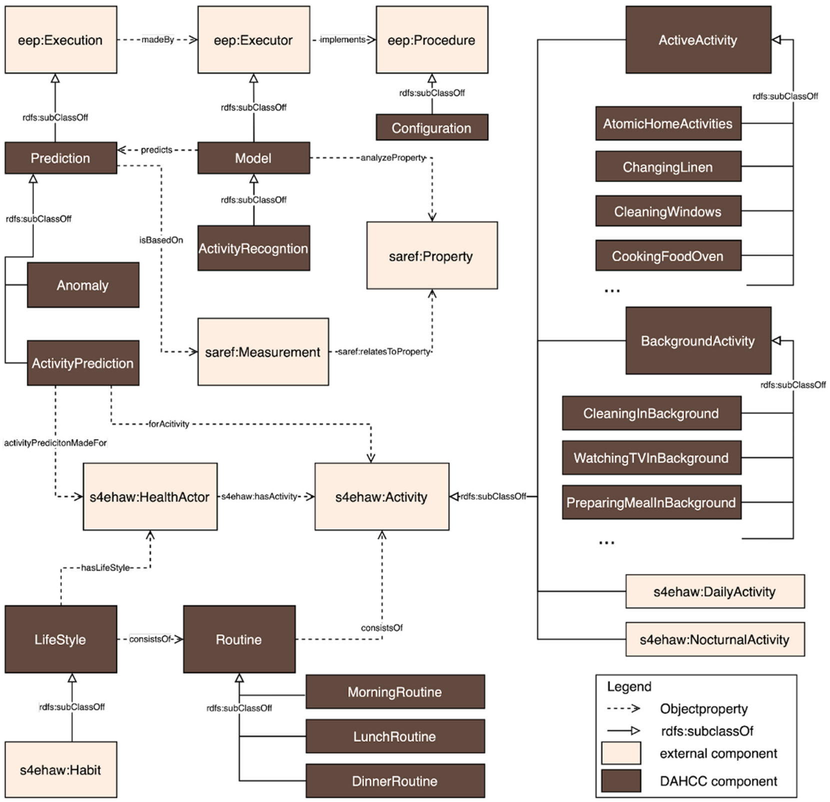


Fig. 1. Overview of the DAHCC Activity Recognition sub ontology

environment for IoT services and smart living. A wide range of IoT technologies are deployed, and the set-up allows to add new devices using technical corridors, hollow floors and ceilings[7].

In total, 42 participants were invited to the Homelab and were asked to either perform their daily or nightly routines. Both the Homelab and the participants were equipped with a large number of sensors which try to capture the participant's activities. The data collected for each participant, as well as the daily life annotated activities, are made available at https://dahcc.idlab.ugent.be/dataset.html.

4.1 Data Collection Setup

To derive the daily and nightly activities, a data platform was designed to capture and store the sensor readings and participant annotations. An overview of this platform is provided in Fig. 6. The rooms within the Homelab were

---

[7] https://www.ugent.be/ea/idlab/en/research/research-infrastructure/homelab.htm.

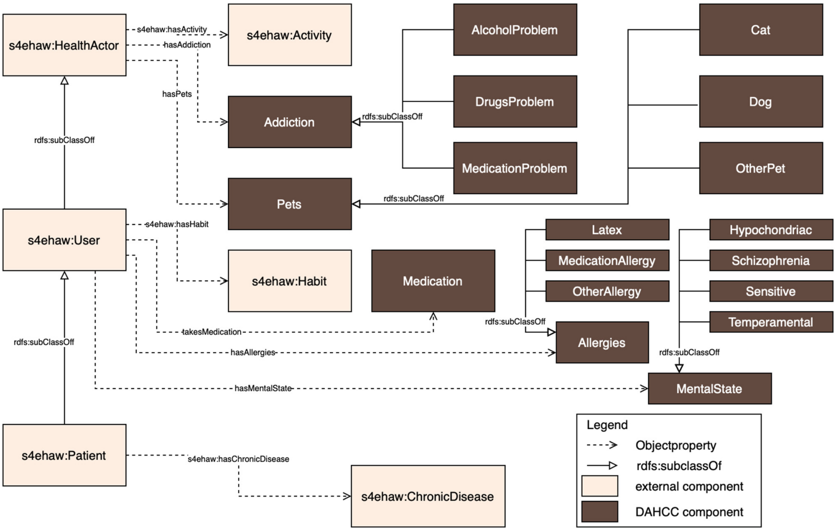


Fig. 2. Overview of the DAHCC Monitored Person sub ontology

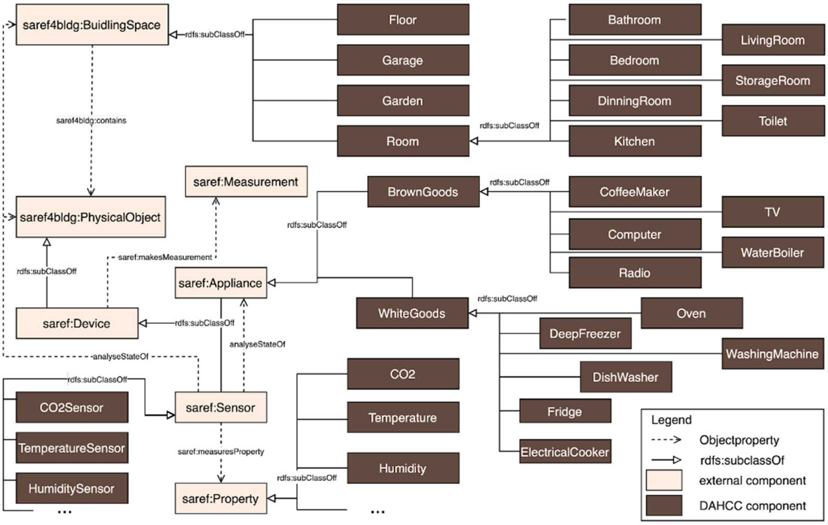


Fig. 3. Overview of the DAHCC Sensors and Actuators sub ontology

also equipped with many contextual sensors. People Counters and the AQURA Indoor localization system of Televic Healthcare[8] were used to derive the indoor location of the participants. Door contact sensors were installed on every door,

---

[8] https://www.televic-healthcare.com/en/solutions/wireless-nurse-call-systems/wireless-localisation-tags.

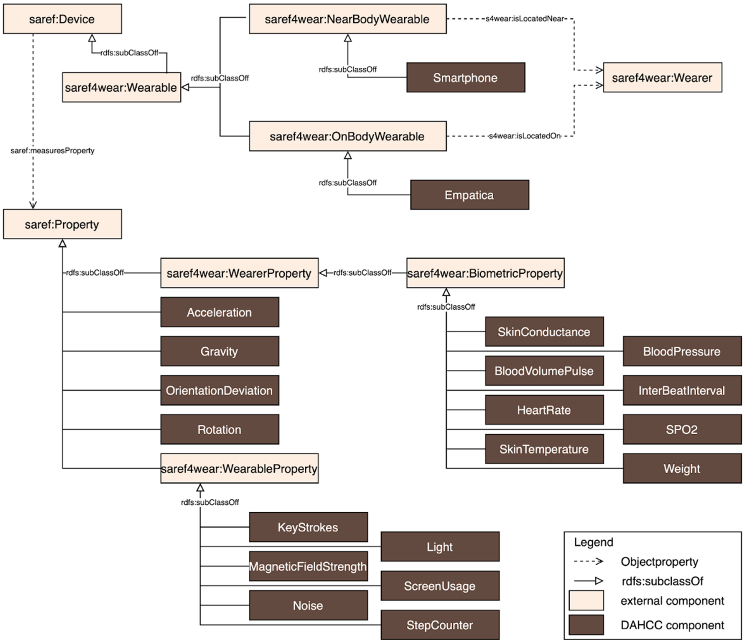


Fig. 4. Overview of the DAHCC Sensors and Wearables sub ontology

window and cabinets in the kitchen the grab their open and close states. Velbus sensors[9] were used to control and monitor the lights, indoor temperature, opening or closing of the blinds and the energy consumption of all appliances. A specifically designed water running sensor was used to detect when water was being used from the faucets in the bathroom and kitchen. At last, several rooms were equipped with a CO2, Humidity and Loudness sensor. All these Homelab sensors were integrated using the DYAMAND platform [15]. DYAMAND collects the data from these sensors and provides it in a JSON format to the data lake.

A web application was designed that allows participants to (a) enter the activities they perform as part of a routine, and (b) indicate when they start/stop a specific activity. The app was used during the whole data collection to allow participants to annotate the actions they are performing. Every time a participant interacted with the application (when a start, end or cancel button was pressed), a timestamp together with the performed action was sent to a log file on a cloud document store. This log file was later on analysed to derive the annotations.

Together with this annotation app, each participant was asked to install two additional smartphone applications: (a) the streaming application to collect data from a wearable (Empatica E4[10]) and smartphone sensors and send them to the

---

[9] https://www.velbus.eu.

[10] https://www.empatica.com.

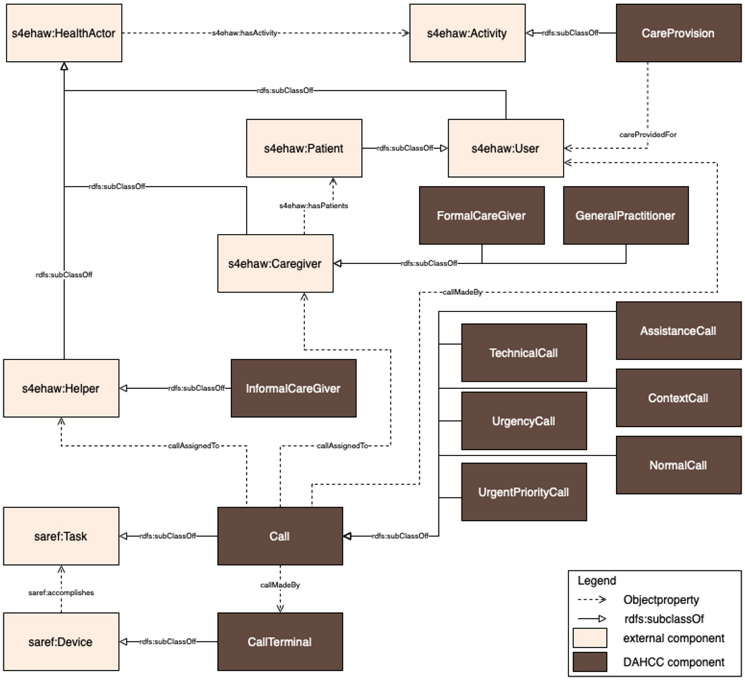


Fig. 5. Overview of the DAHCC Caregivers sub ontology

data lake, and (b) the Sleep as Android application[11] (to track sleep during the night protocol). Besides the wearable device, the blood pressure, body weight, body temperature and spO2 biomedical parameters were measured at the start of the day if the participant gave their approval inside the informed consent.

4.2 Collected Data

In total, 31 "day in life" participants and 12 "night" participants enrolled in this data collection campaign and annotated, on average, 70.7 activities using the mobile annotation app. On average, more than 1 gigabyte of data was collected for every participant. Before making the gathered data publicly available, the data samples were anonymized by erasing the date information within the timestamps of the sensor values (the original times were kept). Also, the participant numbers were randomised, such that participant 1 isn't the first participant who gathered data within this data collection campaign.

---

[11] https://sleep.urbandroid.org.

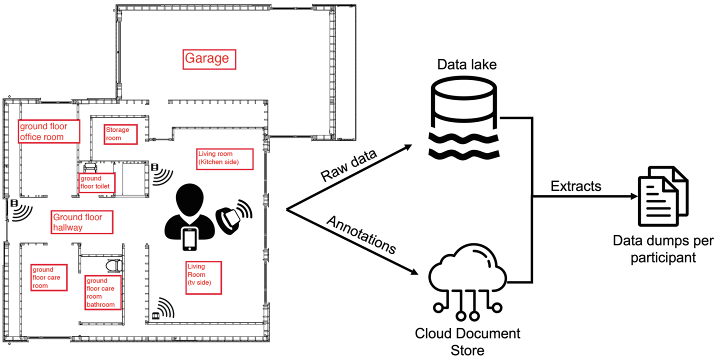


Fig. 6. Overview of the DAHCC Homelab data collection campaign. Data captured from different sensors and wearables was sent to the data lake. Data dumps for each participant with sensor data and annotating labels are generated and made available.

5 DAHCC KG

The described dataset in Sect. 4 was transformed based on the DAHCC ontology into a KG linking all information together. To create this KG, we performed to following steps:

- In a first step, we mapped the Homelab floor plan and location of all sensors and actuators based on the DAHCC Sensor and Actuators sub ontology. We also semantically defined all the major appliances and provided the link to those sensors that measure their energy consumption. The semantic representation of the Homelab is made available as an additional resource[12].
- Secondly, we also mapped the used Empatica E4 wearable and other biomedical devices using the Sensors and Wearables sub ontology. Again, we made this resource available for future reference.
- In a third step, all the data from the data collection participants were transformed into a semantic representation using a Python script. This script maps each sample to a participant-specific URI identifier and links it to the concept described in the Sensors and Actuators and Sensors and Wearables sub ontologies.
- A similar Python script was created to map the participants' annotations onto the concepts of the Activity Recognition sub ontology.

For each participant, the output of these scripts generated NTRIPLE files, which were combined and gzipped to reduce the KG file size. Those individual KG files, per participant, were also made available at https://dahcc.idlab.ugent.be/dataset.html.

---

[12] https://github.com/predict-idlab/DAHCC-Sources.

6 DAHCC Semantic Applications

Inspired by the available DAHCC KG, we created three applications which can be used to derive new knowledge from or infer useful results for healthcare professionals. A first application defines how the available domain knowledge within the DAHCC ontology can be used to generate more advanced events using reasoning. A second application describes how semantic rule mining based on the inferred knowledge can be performed. A third application shows how such a rule can be incorporated into a semantic stream reasoning engine. Implementations and examples on how to use each of these applications are provided in a GitHub repository[13].

6.1 Semantic Higher-Order Events Generation

The DAHCC KG defines the sensor data and corresponding metric information. Due to all the available knowledge, we can reduce this KG by transforming groups of sensor observations into more relevant events happening inside the Homelab. Most sensors measured the state of an appliance or object within a certain room, e.g. the Velbus energy sensors measure the energy consumption of the cooking top in the kitchen. Instead of storing only the data observations in a semantic format, the usages of appliances and objects were also inferred based on the sensor values inside the KG (Fig. 7).

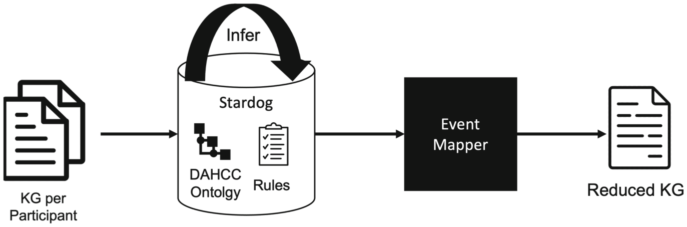


Fig. 7. Overview of the Inferred Knowledge generation. This generator was adapted from https://www.stardog.com/labs/blog/stream-reasoning-with-stardog/

In this approach, the semantic observation samples are loaded within a Stardog[14] database together with the DAHCC ontology and some generic, predefined rules. Reasoning-on-query is then executed on these observation samples within the Stardog database to get the inferred and derived events. For each defined rule, a specific event is created when inferred (e.g. the Start command rule creates an on state event for a specific appliance).

Next to the action and corresponding appliance or physical object related to the occurring action, both the time when this event happens and the participant

---

[13] https://github.com/predict-idlab/DAHCC-Sources/tree/main/Applications.

[14] https://www.stardog.com.

who executes the action are stored. The annotated activities, the inferred events and sensor observations were all combined within a new KG which now only contains events instead of a large number of observations.

Within this application, the DAHCC ontology is used to deliver additional context information to enrich the raw data. Since the specific Homelab appliances, rooms and sensors are described by such DAHCC components, rules could be designed to combine the data and metadata to generate new insights. Those defined rules are due to the use of the DAHCC ontology generically and are easy to interpret.

The creation of higher-order events is needed in a healthcare setting to reduce the number of raw data samples that have to be monitored by a healthcare professional. Due to the available ontology and the metadata described by this ontology, simplifications of the data can be provided to create more interpretable information about what is going on in a smart home.

6.2 Deriving Rules for Shower Events

The previous application generates more advanced events to monitor the behaviour of a person in an ambient house setting. Additionally, it would also be beneficial to detect the lifestyle activities performed by these people as those activities define the current behaviour and state of the person. Based on the available data and metadata, and to ensure the detection method is interpretable, a semantic rule mining application was designed to derive lifestyle rules based on the semantic events generated by the previous application. The INK rule miner [23] was used to derive task-specific rules for one group of activities (e.g. shower events) compared to all other events. This rule miner is based on the INK representation to embed the KG but performs a task-specific rule mining operation based on Bayesian rule set [26]. The task here is to find semantic rules which discriminate against one type of activity as best as possible regarding precision (defining how many predicted rule outcomes were relevant) and recall (defining how many relevant triggered rule outcomes were retrieved). Table 2 shows examples of relevant rules found for the Shower, Toilet and Washing Hands activities. For all tasks, a highly imbalanced set of labels has been provided and the INK rule mining parameters were set to mine the most precise rules. Therefore, the recall scores are significantly lower than the precision scores for the obtained rules.

The DAHCC ontology delivers in this application the additional knowledge to mine more insightful rules. Without the DAHCC ontology, rules will be less generic and it will be harder to mine rules based on related data observations (e.g. observations from similar sensors or observations made within the same room). Mining rules automatically is also needed to create an interpretable, but adaptive healthcare monitoring system where a human only has to verify but not create the rules.

Table 2. Results of the performed rule mining operation on the dataset

Event	Rule	Precision	Recall
Shower	hasEvent.kitchen.Temperature > 21.75 and hasEvent.bathroom.Loudness_mean > 46.83 and NOT hasEvent.personIn§kitchen	0.9411	0.6153
Toilet	hasEvent.LightSwitchOnIn§toilet1 and NOT hasEvent.personIn§kitchen	0.8175	0.5734
WashingHands	hasEvent.LightSwitchOnIn§toilet1 and hasEvent.using§waterpump	0.8600	0.3385

6.3 Semantic Stream Reasoning

Mining rules based on semantic events deliver useful insights into the daily life pattern of a person. Rule-based systems are frequently used as connected care applications as they are both reliable and interpretable. We created a semantic stream reasoning application to show how DAHCC can be used for such a technique.

An overview of this semantic stream Reasoning unit is provided in Fig. 8. First, the raw data samples are mapped in a semantic format one by one. Next, those semantic observations are fed to a C-SPARQL [2] engine. C-SPARQL is used here in combination with an RSP-Service[15]. This combination of RSP and C-SPARQL makes it possible to dynamically load query rules to detect the lifestyle activities of a person. In the last step, the query is performed on a window of obtained semantic events and when a result can be inferred, this semantic result is sent to a monitoring application.

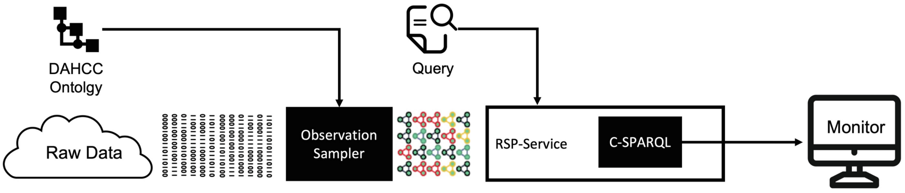


Fig. 8. Overview of the semantic stream reasoning setup

An instantiated application where we try to find shower activities within a semantic data stream is made available. Again the DAHCC components described the semantic observations, on which the shower query rule is defined. Multiple sensor values to determine the location of the person and the humidity

---

[15] https://github.com/streamreasoning/rsp-services.

sensor values within the bathroom are combined in order to generate a rule-based prediction. The whole setup and the defined rule-based predictions are also semantically described using the DAHCC ontology.

In the healthcare sector, semantic stream reasoning can be used to monitor a patient with specific care needs. Cases exist where based a smart room adapts to the needs of a patient suffering a concussion [6]. Using the DAHCC ontology different alert levels and priorities can be automatically defined and monitored based on the patient's care needs and the available sensors and actuators inside a smart home.

7 Discussion and Conclusion

In this work, we presented DAHCC, a combined resource which provides healthcare knowledge in numerous settings using a maintained ontology and a large dataset to build and create semantic connected care applications. All the resources, the resource creation files and example semantic connected care applications files are made open source.

The applicability of the ontology is evaluated by transforming the dataset into a semantic format, resulting in a KG. The construction of these KG files was created using a script because all the original raw dataset files had a similar structure. Techniques exist to provide a more user-friendly and standardized way to transform such data files into a semantic representation. The applicability and overview of these existing techniques, their benefits and drawbacks, were left out of scope in this research.

The open-sourced KG is based on all raw sensor input from a smart lab environment. Therefore, the KG shows only one part of the available DAHCC ontology. The goal of the DAHCC ontology is that additional instances such as designed artificial intelligence models or patient-specific information are also made available in such a KG. The design of such an artificial intelligence model is kept to a minimum and is part of future work. Patient-specific information was not incorporated in the dataset as anonymization of the participants was required to make it publicly available.

At last, the ontology itself was built by taking into account the input from different people within the healthcare domain. While the current setting was mainly focused on how ambient living and connected care systems can be combined, the ontology itself is defined and constructed to be further extended or adapted when new information or new use cases become available. Making the ontology open-source makes it possible for other researchers to further extend it.

Resource Availability Statement: The DAHCC ontology, datasets and KG are available online from https://dahcc.idlab.ugent.be. The source code for the applications is available on Github at https://github.com/predict-idlab/DAHCC-Sources.

Acknowledgement. Bram Steenwinckel (1SA0219N) is funded by a strategic base research Grant of the Fund for Scientific Research Flanders (FWO). This research is part of the imec.ICON project PROTEGO (HBC.2019.2812), co-funded by imec, VLAIO, Televic, Amaron, Z-Plus and ML2Grow.

References

1. Aldahiri, A., Alrashed, B., Hussain, W.: Trends in using IoT with machine learning in health prediction system. Forecasting **3**(1), 181–206 (2021)
2. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: C-sparql: Sparql for continuous querying. In: Proceedings of the 18th International Conference on World Wide Web, pp. 1061–1062 (2009)
3. Chiang, L.C., Chen, W.C., Dai, Y.T., Ho, Y.L.: The effectiveness of telehealth care on caregiver burden, mastery of stress, and family function among family caregivers of heart failure patients: a quasi-experimental study. Int. J. Nurs. Stud. **49**(10), 1230–1242 (2012)
4. Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare. Future Healthc. J. **6**(2), 94 (2019)
5. De Backere, F.: Towards a social and context-aware multi-sensor fall detection and risk assessment platform. Comput. Biol. Med. **64**, 307–320 (2015)
6. De Brouwer, M., Ongenae, F., Bonte, P., De Turck, F.: Towards a cascading reasoning framework to support responsive ambient-intelligent healthcare interventions. Sensors **18**(10), 3514 (2018)
7. De Brouwer, M., et al.: mBrain: towards the continuous follow-up and headache classification of primary headache disorder patients. BMC Med. Inform. Decis. Mak. **22**(1), 1–34 (2022)
8. Elhadj, H.B., Sallabi, F., Henaien, A., Chaari, L., Shuaib, K., Al Thawadi, M.: Do-care: a dynamic ontology reasoning based healthcare monitoring system. Futur. Gener. Comput. Syst. **118**, 417–431 (2021)
9. Ganesan, A., Paul, A., Seo, H.: Elderly people activity recognition in smart grid monitoring environment. Math. Probl. Eng. 2022 (2022)
10. Hofman, J., La Manna, V.P., Muylaert, J.: Measuring and modeling air quality in smart cities (2021)
11. Jin, W., Kim, D.H.: Design and implementation of e-health system based on semantic sensor network using IETF YANG. Sensors **18**(2), 629 (2018)
12. Jørgensen, D.B., Hallenborg, K., Demazeau, Y.: Patient centric ontology for telehealth domain. In: Geissbühler, A., Demongeot, J., Mokhtari, M., Abdulrazak, B., Aloulou, H. (eds.) ICOST 2015. LNCS, vol. 9102, pp. 244–255. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19312-0_20
13. Mamdiwar, S.D., Shakruwala, Z., Chadha, U., Srinivasan, K., Chang, C.Y., et al.: Recent advances on IoT-assisted wearable sensor systems for healthcare monitoring. Biosensors **11**(10), 372 (2021)
14. Moreira, J., Pires, L.F., van Sinderen, M., Daniele, L., Girod-Genet, M.: Saref4health: towards IoT standard-based ontology-driven cardiac e-health systems. Appl. Ontol. **15**(3), 385–410 (2020)
15. Nelis, J., Verschueren, T., Verslype, D., Develder, C.: Dyamand: dynamic, adaptive management of networks and devices. In: 37th Annual IEEE Conference on Local Computer Networks, pp. 192–195. IEEE (2012)
16. Ongenae, F., et al.: An ontology co-design method for the co-creation of a continuous care ontology. Appl. Ontol. **9**(1), 27–64 (2014)

17. Ongenae, F., et al.: Participatory design of a continuous care ontology : towards a user-driven ontology engineering methodology. In: KEOD 2011: Proceedings of the International Conference on Knowledge Engineering and Ontology Development, pp. 81–90. INSTICC (2011)
18. Peng, C., Goswami, P.: Meaningful integration of data from heterogeneous health services and home environment based on ontology. Sensors **19**(8), 1747 (2019)
19. Reda, R., Piccinini, F., Carbonaro, A.: Towards consistent data representation in the IoT healthcare landscape. In: Proceedings of the 2018 International Conference on Digital Health, pp. 5–10 (2018)
20. Rhayem, A., Mhiri, M.B.A., Gargouri, F.: HealthIoT ontology for data semantic representation and interpretation obtained from medical connected objects. In: 14th International Conference on Computer Systems and Applications, pp. 1470–1477. IEEE (2017)
21. Schinköthe, T., et al.: A web-and app-based connected care solution for COVID-19 in-and outpatient care: qualitative study and application development. JMIR Pub. Health Surveill. **6**(2), e19033 (2020)
22. Sondes, T., Elhadj, H.B., Chaari, L.: An ontology-based healthcare monitoring system in the internet of things. In: 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), pp. 319–324. IEEE (2019)
23. Steenwinckel, B., Pieter, B., De Turck, F., Femke, O.: Ink: Knowledge graph representation for efficient and performant rule mining. Semant. Web J. (2022)
24. Tiwari, S., Abraham, A.: Semantic assessment of smart healthcare ontology. Int. J. Web Inf. Syst. **16**, 475–491 (2020)
25. Tun, S.Y.Y., Madanian, S., Mirza, F.: Internet of things (IoT) applications for elderly care: a reflective review. Aging Clin. Exp. Res. **33**(4), 855–867 (2021)
26. Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., MacNeille, P.: A Bayesian framework for learning rule sets for interpretable classification. J. Mach. Learn. Res. **18**(1), 2357–2393 (2017)

Health Data Semantics: Exploring Requirements for Sustainable Health Systems

Mate Bestek[1,2](✉) and Erik Grönvall[1]

[1] IT University of Copenhagen, Copenhagen, Denmark
`{mbes,erig}@itu.dk`
[2] 3fs d.o.o., Kranj, Slovenia
`mate.bestek@3fs.cloud`
`http://www.itu.dk,http://3fs.cloud`

Abstract. Health data cannot easily move from one healthcare provider to another, let alone between different countries. There is a de facto lack of health data interoperability in many healthcare systems preventing mobility and sharing of health data. Hitherto the literature has mainly considered technical aspects and challenges related to interoperability, but this paper will foremost explore semantic aspects. Using two cases of exploring data interoperability and the democratisation of health data as a backdrop, this paper presents existing challenges preventing sustainable health systems by focusing on semantic aspects of data interoperability in healthcare. A key finding is that data interoperability is possible and standardised data semantics are key resources to support data access within or across systems, for example between countries. The democratisation of health data, allowing different health systems to interact with each other, requires management of the shared semantic resources and the governance structures defining them. Domain experts, like medical doctors, can play an important role in co-designing and managing such shared, semantic resources.

Keywords: Health Data Semantics · Semantic interoperability · Democratisation · Co-Design · Human-Data Interaction

1 Introduction

It is not hard to imagine a healthcare scenario in which a person becomes involved in an accident while travelling abroad. Let us take a person named Mark who travels to Denmark. It is not his first time there, but this time Mark gets hit by a car and loses consciousness due to a blow to his head. Mark is brought unconscious to the hospital where a physician orders a magnetic resonance scan to understand if there are internal injuries to the brain. But what if Mark has had previous accidents and has a metal plate implanted in his head? The treating

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 361–375, 2023.
https://doi.org/10.1007/978-3-031-34586-9_24

physician does not know Mark, nor does he have access to his medical record. In such situations and any other situation where medical decisions need to be made, access to information from existing medical records is fundamental. One could not even get a medication dispensed abroad due to issues of differing dosages, compounds, and packaging of medications used in different countries - in spite of many project initiatives funded by the EU, e.g. epSOS, MyHealth@EU, that have greatly influenced our previous work [1,2]. At this point, one could say that such cross-border scenarios do not happen often. But the Covid pandemic forced us to lift priority for such scenarios. The European countries, for example, were collectively able to quickly implement a standard in the form of a Covid passport that enabled travel between countries. However, as we explain in Sect. 4, no health data has been transferred between countries with the European implementation of the so-called Covid passports. But to help Mark transfer his electronic medical record to Denmark, we need to transfer medical data and solve much more complex underlying issues. As we will also learn in this paper, these underlying issues are also preventing health data flow within countries or even between neighbouring hospital wards (naturally depending on how each country and region manage their health systems). The above-mentioned issues are categorised on different levels of interoperability, for example technical, semantic, organisational, and legal. These levels of interoperability all play a role in establishing the capability of data exchange in which data is understood in the same way by all parties irrelevant of location or language used [15]. Interoperability problems are difficult to solve due to the complexity of health systems and the large number of stakeholders involved at different levels and with different roles. As a result, health data are scattered across health systems leading to health data being an underused resource that cannot flow where needed since manual health data transfer between systems is expensive, highly labour intensive, and prone to semantic inaccuracy [38]. Figure 1 shows an example of semantic inaccuracy on the blood oxygen level data stored in two different databases. What we see in the example from Fig. 1 and from literature (e.g. [38]), is that health data are defined in different ways which are usually not compatible without additional data work. One solution to reduce the need for such additional data work is to define and store data in the same way in both databases. That is, to establish shared semantic resources as standardised definitions of health data to ensure a standardised interpretation and understanding of data, regardless of where, when and by whom the data is used [2]. Such shared semantic resources are at the core of solving data interoperability. However, such approaches heavily influence how information systems are designed and developed in the first place and thus impose constraints on e.g. EHR system providers. But in order to reduce the semantic difference between data coming from different systems in a highly complex domain such as healthcare without compromising on the completeness of individual patient's health data, we have pursued such solutions. To define how to store data, OpenEHR is an existing architecture that supports semantic work - the creation of shared semantic resources in healthcare [22]. OpenEHR is a blueprint for how to set up a technical infrastructure for storing and managing

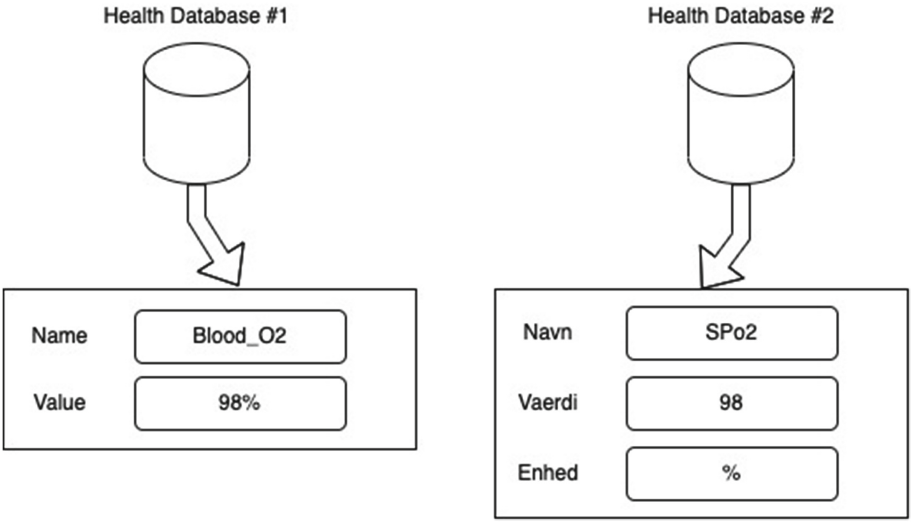


Fig. 1. We see differing structures and semantics for a data structure that captures a simple measurement of blood oxygen levels in the blood. The database on the left only uses two data fields to store the measurement while the one on the right uses three data fields. In addition, the names of the fields differ due to the different languages used.

electronic health records [2]. OpenEHR "consists of open specifications, clinical models and software that can be used to create standards and develop information and interoperability solutions for healthcare" – [26]. Previous research provides insights on how to sustainably establish semantic work by promoting the process of democratisation of health data semantic resources [1,2] or sharing of health data semantics [40]. Democratisation in the context of data has been defined by Samarasinghe et al. as "*an ongoing process of enabling digital data access to both technical and non-technical users to understand, find, access, use, interact and share appropriate data...*" - [33]. Influenced by this data democratisation definition by Samarasinghe et al., we can define health data semantics democratisation as an ongoing process of enabling access and sharing of health data semantics. The democratisation of health data semantics can be seen as an alternative approach to the currently prevailing lock-in-based business strategy. In lock-in-based business strategy software companies hide the semantic resources that define the data in their systems. The end goal of such behaviour is to force the dependency of end users on the companies for the firms to ensure themselves long-term business engagements [1]. We do understand, however, that protecting intellectual property rights might be one of the factors influencing the need for such business models. Using two case studies from Slovenia, this paper focuses on understanding semantic work as a prerequisite for achieving democratised or shared semantic resources. These semantic resources enable a shared understanding of data that can then flow where needed. That is, shared semantic resources enable data interoperability and data flow not only within complex health systems environments with many different stakeholders with different roles in specific countries or regions but also between health systems of different countries. In this paper, we try to point out the need to recognise semantic work as a crucial prerequisite of shared health data semantic resources which can support more sustainable health systems in the future. In this paper, we point to existing very important legal mechanisms that the EU is currently rolling out that could help achieve that goal. The rest of the paper continues as follows: in Sect. 2 we introduce important concepts that we use in the paper; in Sect. 3 we describe how our data was obtained which we describe in Sect. 4. In Sect. 5 we analyse and explore the empirical data and what we can learn from it. Finally, we wrap up the paper with a discussion and contributions in Sect. 6.

2 Related Work

Due to the importance of understanding semantic work we try to align with existing research work to inform our understanding of semantic work. This includes aligning with work on co-design that can provide methods used as part of semantic work. In addition, it includes aligning with research on human-data interaction where interesting insights can help position the importance of semantic work. After bringing forward these two sources of inspiration for our research, we shortly review the existing state of standardisation and legislation activities in the EU to support our discussion on how to establish a more long-term sustainable and scalable semantic work. This first includes the Covid-19 passport activities that have resulted in an implementation of a standardised solution across the EU member states. This work is important as it provides evidence of the successful tackling of interoperability problems between countries. However, as it lacks focus on semantic resources and semantic work, we also provide here a short overview of some of the relevant EU legislation that has lately been adopted or is still in the process of being adopted, as such legal frameworks may help with achieving a sustainable and scalable semantic work in the future.

2.1 Co-design and Participatory Design in Health Care

Co-design and Participatory Design (PD) are often treated as synonyms and described as an approach to design together with end-users and other stakeholders [12,13,34]. With its roots in work on knowledge development, and early projects like Utopia, Participatory Design (PD) first emerged as an effort to include workers in the design of workplace computer systems and later matured into an approach to design with different stakeholders [10,36]. The healthcare sector has from early PD projects to more contemporary co-design projects been the context of multiple collaborative design projects [3,23,39]. Relevant to this paper, PD and co-design have been used to design collaborative health systems that support different professional and informal carers in their collaboration [5,6]. These papers have for example highlighted the importance of different stakeholders collaborating, also across organisational boundaries, to support a positive care trajectory. PD has for example been used to involve people with dementia and their carers [37], and co-design has been used to promote knowledge creation among stakeholders [27]. Recent work by Grönvall et al. has looked into how social workers can design training materials for their clients on a collaborative platform disregarding which care organisation or municipality they are employed in [24]. To sum up, the collaborative design process is not new in healthcare and has been used to improve learning, foreground knowledge within cross-organisational communities and develop care systems and technologies.

2.2 Human-Data Interaction

In recent years an increased focus on Human-Data Interaction has emerged as a complement to the broader field of Human-Computer Interaction [19,29,41]. Human Data interaction emphasises how people work with, and generate, data rather than the interactions people do with interfaces inserted between them and the data sources. Within a large part of the healthcare sector data (both individual patient health data and Big data) is seen as an enabler for effective and quality care, and Electronic Healthcare Records (EHR) are used in many parts of the world [7]. Within healthcare, HDI has been used to explore how we can for example make better use of data [11] allowing for example healthcare professionals to engage in data-work [4]. Bossen et al. describe how data is not always readily available and may have to be combined, analysed, and contextualised to be valuable; they define data work as “any human activity related to creating, collecting, managing, curating, analysing, interpreting, and communicating data” [7]. Work related to HDI outside of the healthcare domain has also explored how experts in non-IT domains can turn data into objects of design and explore how to use data in new and innovative ways [34]. Seidelin et al. discuss for example scenarios where it is useful to explore HDI from a multi-stakeholder and multi-organisational perspective [35]. Seidelin et al. also recognise that in such scenarios it can be fruitful to involve data users in the design of how data is stored, used and analysed and they identify co-design as a way to meaningfully involve the different actors [34].

2.3 EU Digital Covid-19 Certificate as a Cross-Border Standard Implementation

During the Covid-19 crisis, a transnational need emerged to allow Covid-tests (e.g. Antigen and PSR) performed by a health institution in one country to be valid and verifiable in other countries. In response to that need, the European Union developed a digital documentation standard for Covid-tests and vaccine certificates valid across the European Union. Key features of the certificate are according to the European Union [14], that it works for both in digital and/or paper format, it is based on a QR code, it is free of charge and available in both the national and English language. It’s also secure and valid in all EC countries. The functionality is built around the above-mentioned QR code that can be scanned at for example airports, shopping centres, restaurants, and bars.

The Covid passport app (and its paper-based alternative) is an interesting solution in terms of interoperability and trans-national mobility of health data and certificates that prove that a person has received a Covid-19 vaccine or been tested negative no matter where in the EU (or world if they accept the standard) the verification is performed. The generated QR code contains the signature of the local health provider in a specific country that has issued the certificate. Based on an EC-wide common format and an EC-level gateway the Covid passport validity can be verified without sending an individual's actual health data between countries.

The above example is interesting as it enables a cross-national validation of medical data. However, it does not work in the case of medical records where actual health data has to be moved or accessible from different nations or between providers. A health record is also far more complex and contains far more data that could hardly fit into a QR code. In addition, this case is important for this paper because it shows that a cross-national solution is possible in the EU. This means that existing legal frameworks and technical infrastructure can be used to achieve such goals. However, the crucial aspect of this case that is important for this paper is the complete lack of focus on health data which is never actually carried abroad. For this, there is no need to handle semantic resources which greatly simplified the overall design and implementation.

2.4 Standardisation and the European Union Health Data Spaces Regulation

The European Union (EU) Strategy on standardisation "*aims to strengthen the EU's global competitiveness, to enable a resilient, green and digital economy and to enshrine democratic values in technology applications*" - [16]. The strategy only recognises standards that are developed by a recognised European Standards Organisation (CEN, CENELEC, or ETSI) upon request from the European Commission. Then, once accepted, such standards become part of EU law and are provided to manufacturers across the European Single Market. Initiatives like the HSBoosterEU project have been initiated to offer standardisation services to existing projects funded by the EU to pursue the implementation of the EU standardisation strategy in practice. However, the strategy and the HSBoosterEU project represent initiatives that mostly fall under legislative and policy matters, and through standards also technical aspects of health data sharing (e.g. epSOS project [28]), but do not address the more complex health data semantics. A step in that direction can be seen in the latest European Health Data Spaces regulation [18].

The European Union has presented the Health Data Spaces Regulation proposal to better use health data. The proposal (1) supports individuals to take control of their health data, (2) supports the use of health data for better healthcare delivery, better research, innovation and policy making, and (3) enables the EU to make full use of the potential offered by a safe and secure exchange, use and reuse of health data [18]. The European Health Data Space is a health-specific ecosystem comprised of rules, common standards and practices, infrastructures and a governance framework with the goal, among others, to foster a genuine single market for electronic health record systems. Since trust is a fundamental enabler for the success of the European Health Data Spaces, it will provide a trustworthy setting for secure access to, and processing of, a wide range of health data. To achieve that, the European Health Data Spaces Regulation build on the General Data Protection Regulation (GDPR) [31], the Data Governance Act [32], draft Data Act [17], and the Network and Information Systems Directive [20].

3 Methods and Materials

This article is based on empirical work involving OpenEHR and different healthcare projects. In particular, we use data from a workgroup (WG) in Slovenia that tried to consolidate semantics across several national e-health projects which involved key stakeholders such as doctors and the government. The data includes primarily documents and communication exchanges which have been made anonymous. For this paper, inductive thematic analysis was used, as the themes identified are strongly related to the data itself and do not come from a theoretical framework [8]. We convey the themes in the following section in the form of two case studies that represent empirical data in this paper. These two case studies can be seen as the contrast to the already introduced case of the Covid-19 passports in Sect. 2 where we learned that in reality no health data was exchanged between EU countries. We have chosen the Covid-19 passport as a contrasting case because it is a public and well-known example of a standardised solution that enables sharing of health data, albeit only implicitly utilizing digital signatures, across national borders and health systems. With this contrast, we try to point to the importance of semantic work as was done within the WG. We see such semantic work as the basis for achieving shared or democratised semantic resources.

4 Empirical Data: Two Case Studies of Semantic Work

We will now outline two real-life cases to help us in our work. The two examples come from the exploration of health data interoperability through the development of semantic resources by the aforementioned National e-Health Programme Working Group (WG). We contrast these two cases with the aforementioned example of the Covid-19 passport where the certificate allows, for example, a tourist from one European country to visit another using a Covid test certificate issued at home.

4.1 Scenario 1: The Asthma Questionnaire

The work group of the national e-Health programme (WG) worked on different types of semantic resources during its existence. The WG worked on a specific sub-project of the national e-Health programme that focused on creating national datasets for the primary level of the health system so that the Ministry of Health could better track the health status of the population.

Clinically validated questionnaires are one of the most important semantic resources in healthcare. These are used as medical tools which are often developed by pharmaceutical companies which raises ownership issues and issues with intellectual property if misused. The WG encountered such questionnaires, like the asthma health status questionnaire, that is used to assess patient asthma status. Since the WG worked on a specific project that included a limited number of use cases on the national level, the asthma questionnaire could not just be used as part of the national semantic resource set. To be able to do that, a legal agreement had to be established with the owner of the intellectual property over the questionnaire, being a large international pharmaceutical company. One of the participants of the WG - a medical doctor - offered to obtain permission from the pharmaceutical company that would allow the use of the questionnaire in the national project. The permission was obtained in a form of a contract that defined the rules of use for the questionnaire at hand.

We consider this case important for this paper because it depicts specific semantic work. It points out a specific semantic resource - the asthma questionnaire - that has specific requirements to be governed properly as a shared semantic resource. We can also learn that the direct involvement of e.g. the Ministry of Health was not needed or required in our case, even if it dealt with the national governance of semantic resources that are under the protection of intellectual property legislation. At the same time, this case shows how members of the WG were able to find a solution non the less. However, the solution can be seen to only be partial and could not scale to include potential future use cases or projects. Future projects could perhaps not be approved by the pharmaceutical company which could also pose payment of high fees for the rights of use.

4.2 Scenario 2: Doctor-Created Data Structures

The WG engaged with several other semantic resources including clinical concepts like blood pressure, blood glucose and disease diagnosis. Each such clinical concept was discussed in terms of what additional data is needed to fully capture the context within which a measurement is made or a data point captured. This was done by discussing a set of support data for each clinical concept to achieve a sound structural definition of the concept. In addition, each support data element could include values from existing medical terminologies. Therefore, the WG semantic work resulted in defining precise data structures and mappings to different terminologies for the clinical concepts that were important for the project at hand. At this point, it is important to stress that the members of the WG, by referencing the OpenEHR approach, were able to obtain existing data structures and mappings to terminologies for several of the clinical concepts they worked on. The global OpenEHR community works on such concept definitions that can be freely used. However, despite the convenience of being able to learn from and even reuse some of the existing concept definitions, the semantic work also included making changes to existing concept definitions or even creating new ones. In some cases, this also included work on reducing the complexity of some data sets and clinical concept definitions. An example of such simplification was a rather complex survey-like questionnaire. The reasoning for the simplification was that simpler semantic resources could be reused to design different more complex semantic resources. In this case, some common elements of the complex questionnaire could become reused in different questionnaires. In addition, simpler semantic resources could also offer a superior user experience for the end users when implemented in technical systems. Another example of doctors working on semantic resources can be seen in their collaboration on a single code from the International Classification of Diseases (ICD). It turned out that different doctors used different codes to identify smokers and non-smokers. One doctor pointed out that he uses the ICD code F17.1 to identify a smoker and Z000 to identify a non-smoker. But the doctor was eager to define a better set of codes if that would be needed. We consider this case important because it introduces additional aspects of the semantic work done by the members of the WG. It depicts how the WG jointly worked on finding consensus on several clinical concepts definitions that included work on data structures, like the overly complex survey-like questionnaire, and mappings to different terminologies, like the codes for identifying smokers and non-smokers. Such work is the essence of semantic work but is currently not recognised by health systems as a permanent activity and a need. This means that in spite of the fact this WG existed and worked on data structures and terminologies, they stopped their work immediately after preparing their results. These results - the semantic resources - were then used to develop a technical information system but unfortunately, the semantic resources have not become part of a national governance initiative of semantic resources that could ensure these semantic resources would be used, reused or improved in future project initiatives.

5 Analysis and Results

We have learned about a successfully implemented and interoperable cross-border Covid-19 certificate solution enabling people to travel between countries based on their vaccination status. However, this implementation did not require any health data from a person's medical record to travel across borders. The EC Covid-19 passport implementation was based on a digital signature that had to be verified and it was then implicit what your vaccination status was. To enable electronic medical records to be transferred between health systems, even between countries, something that would have been useful for Mark from the beginning of the paper, is a much more complicated task. Even if we do not consider complex technological and data security challenges, the very practice of using health data originating in health systems other than your own can be difficult. It requires collaborative work on health data structures and their semantics - what we call semantics work - is not only considered in local contexts but also international health data exchange. Our two empirical cases show that medical doctors can collaborate on defining data structures, vocabularies and their harmonisation. Looking at such semantic work through the lens of human-data interaction and the work of [34], it is beneficial to have multiple stakeholders involved in such activities in the early stages of healthcare system design, where it is defined how data is stored in the first place. Seidelin et al. [34] point out that the active involvement of data users can be of great importance in designing the data structures that will later store the data. Our two empirical cases can thus be seen as instances of co-design that resulted in health data structures and a shared understanding of the meaning of different health data. In our two empirical cases, the co-design process utilised the methodology of OpenEHR to arrive at the designed objects - the health data semantic resources.

Unfortunately, semantic work - despite being crucial for achieving health data semantics democratisation and consequently trustworthy health data interoperability, is usually not long-lived. Such work is mainly temporary and does not scale on a national or international level and therefore often remains on the level of a small-scale project. More importantly, semantic work is often not recognised as a part of the health system as no legislation supports or requires this work to happen. As has been explored in [1,2], it would be desired to have such data interaction work in the early design stages of technical information systems development and it needs to become an ongoing activity of health systems because such work is a needed prerequisite for achieving trustworthy and interoperable health data exchange. As we have learned, human data interaction work benefits from having multiple stakeholders participate in the co-design of data structures. In Europe, the European Commission is one such needed stakeholder that is of crucial importance. We have also learned from our empirical cases that focusing only on semantics is not enough as even in semantic work there are e.g. legal questions to be addressed. Having a European Commission as a stakeholder dealing with such legal issues would gravely impact the implementability, sustainability and scalability of semantic work. As we have learned in Sect. 2, Europe has prepared strategies and legal grounds that support future health data exchange between countries. These strategies and legal frames can help support the collaborative design of health data structures and health data semantics - semantic work - in a more sustainable and scalable way. We observe a link between the semantic work as a need of health systems and existing EU strategies and legislation that have great potential for addressing that need.

In particular, the EU Standardisation Strategy is a mechanism through which new standards can be introduced to all the EU member states' health systems. This means that standardised health data structures and semantics could be adopted through this process and new requirements for the health systems could be created. In addition to establishing a legal and standards framework, the mechanism could also be used for the ongoing creation and dissemination of new semantic resources. In this way, a sustainable work practice for the governance of semantic resources could be achieved.

In addition, the Health Data Spaces Regulation is a legal framework that supports achieving health data democratisation. This means that standards and procedures could be established that would support implementation. In addition, the regulation brings forward data altruism - data voluntarily made available by individuals or companies for the common good. An interesting avenue for the future would be to consider semantics altruism, a term very similar to shared semantic resources. This idea could perhaps be supported in the Data Governance Act, by potentially introducing governance of data semantics resources and the Data Act where interoperability standards are the focus.

6 Discussion and Contributions

We learned in Sect. 2 that successful implementations of cross-border solutions in the field of healthcare are already possible. An example of such a successful implementation is the EC Covid-19 passport implementation from which we learn that technical, organisational and legal interoperability problems can be addressed in a rather straightforward way following a top-down definition of a standard solution. The crucial challenge lies in the semantic interoperability issues. To solve these issues means establishing shared semantic resources. A joint effort is needed for defining and implementing shared semantic resources in all the technical systems as blueprints of how to store data [38]. This is different to what existing standards like the HL7 FHIR [9,21,25] ensure as their primary focus is on integrating technical systems [30]. However, the Covid-19 passport implementation has been designed in a clever way that completely avoids transferring any health data, and therefore there has not been any need to conduct semantic work as a collaborative design effort between stakeholders and across the EU member states. While the current implementation has successfully enabled people to travel between countries through trust about their vaccination status, it is not the approach needed to solve Mark's problem (from Sect. 1) of transferring electronic medical records between countries. Such implementation is highly dependent on establishing shared semantic resources that define the data elements found in typical medical records. Democratising health data semantics requires software engineers and medical professionals to collaborate on the initial design of semantic resources that define how data is to be stored in the technical information systems and co-create the semantic resources. In this paper, we want to put forward semantic work that we observed in our empirical cases as presented in Sect. 4. Such semantic work involved medical doctors and engineers and should be seen as a need in health systems and addressed accordingly. It will be challenging to get a real impact from the semantic work done by for example the work group in our empirical cases if the results are not considered in a wider context, being regional, national or international (e.g. at the EU level). Referring back to the problem of differing approaches to defining semantic resources - it can be solved by democratising health data semantic resources and the supporting semantics work not only on a local level as in our empirical cases but, more importantly, on an international or EU level. The existing EU standardisation and legislation structures can help achieve that. With this, the important work done by medical doctors as briefly presented in our two empirical cases, should not only be recognised as important but should be recognized as a need of health systems. By recognising semantic work as a need we point to a future in which such work can be made sustainable and scalable, also at an international level like across the EU member states, based on utilising existing standardisation, legal approaches and frameworks.

Our goal in this paper is to bring attention to the semantic work that is already being done in the health systems as exemplified in our two empirical cases but is not sufficiently recognised as a crucial work practice and a need in health systems. We argue that the role of semantic work in healthcare must be further understood if we are to achieve not only democratised health data semantics but also interoperable health data and more broadly, more sustainable health systems. Working with international legislation, like the EU healthcare legislation, is one way to put attention to semantic work in healthcare systems. By doing so, we envision a reality for achieving the participation of the European Commission as a crucial stakeholder in the co-design semantic work efforts, but more importantly, we point to a way to achieve a sustainable semantic work practice that would not only exist during specific projects but would become a cross-cutting ongoing semantic work practice as a solution to the problem of democratisation of health data semantics. Perhaps our last pointer towards such a reality can be seen in referencing the HSBoosterEU project that has been set up in a cross-cutting way, spanning projects, to support standardisation. One can imagine a similar endeavour focused on health data semantic resources.

In this paper, we point out the need for doctors to participate in semantic work to define shared/democratised semantic resources. Such semantic work is currently not recognised as a need and is therefore not supported in the current health system. There is tension between how doctors work and what the health system recognises as work. To help ease this tension we call for better recognition of the semantic work done by the doctors that include very fine-grained work on data structures and terminologies that is the essence of achieving shared semantic resources and with this democratised health data. Due to the important latest legislative work at the EU level, such semantic work could potentially become a reality in the near future.

Acknowledgements. None to declare.

References

1. Bestek, M.: Democratizing health data semantics - a commons-based technology-enhanced activity space to support productive work with health data semantics. PhD, IT University of Copenhagen, Copenhagen (2021). https://en.itu.dk/-/media/EN/Research/PhD-Programme/PhD-defences/2022/PhD-Thesis-Temporary-Verstion-Mate-Bestek-pdf.pdf
2. Bestek, M., Grönvall, E., Saad-Sulonen, J.: Commoning Semantic interoperability in healthcare. Int. J. Commons **16**(1), 225–242 (2022). https://doi.org/10.5334/ijc.1157, https://www.thecommonsjournal.org/articles/10.5334/ijc.1157/. 04 Sept 2022
3. Bjerknes, G., Bratteteig, T.: Florence in wonderland: System development with nurses. Comput. Democracy: Scand. Challenge (1987)

4. Bossen, A.U.C.: Data-work in healthcare: the new work ecologies of healthcare infrastructures. In: CSCW, p. 6 (2016)
5. Bossen, C., Christensen, L.R., Grönvall, E., Vestergaard, L.S.: CareCoor: augmenting the coordination of cooperative home care work. Int. J. Med. Inf. **82**(5), e189–199 (2013). https://doi.org/10.1016/j.ijmedinf.2012.10.005. 13 Sept 2022
6. Bossen, C., Grönvall, E.: Collaboration in-between: the care hotel and designing for flexible use. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 1289–1301. CSCW '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2675133.2675243. 13 Sept 2022
7. Bossen, C., Pine, K.H., Cabitza, F., Ellingsen, G., Piras, E.M.: Data work in healthcare: an introduction. Health Inform. J. **25**(3), 465–474 (2019). https://doi.org/10.1177/1460458219864730. SAGE Publications Ltd
8. Boyatzis, R.E.: Transforming qualitative information. SAGE Publications Inc, Case Western Reserve University (1998). https://uk.sagepub.com/en-gb/eur/transforming-qualitative-information/book7714
9. Braunstein, M.L.: Health care in the age of interoperability part 6: the future of FHIR. IEEE Pulse **10**(4), 25–27 (2019). https://doi.org/10.1109/mpuls.2019.2922575. 13 Sept 2022. ISBN: 9783319934136. IEEE
10. Bødker, S., Ehn, P., Sjögren, D., Sundblad, Y.: Cooperative Design - perspectives on 20 years with 'the Scandinavian IT Design Model (2000)
11. Cabitza, F., Locoro, A.: Human-data interaction in healthcare, p. 10
12. Christiansson, J., Grönvall, E., Yndigegn, S.L.: Teaching participatory design using live projects: critical reflections and lessons learnt. In: Proceedings of the 15th Participatory Design Conference: Full Papers - vol. 1, pp. 1–11. PDC '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3210586.3210597. 13 Sept 2022
13. Ciolfi, L., et al.: Articulating co-design in museums: reflections on two participatory processes. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pp. 13–25. ACM, San Francisco California USA (2016). https://doi.org/10.1145/2818048.2819967, https://dl.acm.org/doi/10.1145/2818048.2819967
14. Commission, E.: EU digital covid certificate. https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/safe-covid-19-vaccines-europeans/eu-digital-covid-certificate
15. Commission, E.: Communication from the commission to the european parliament, the council, the European economic and social committee and the committee of the regions European interoperability framework - implementation strategy (2017). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2017:134:FIN
16. Commission, E.: An EU strategy on standardisation setting global standards in support of a resilient, green and digital EU single market (2022). https://ec.europa.eu/docsroom/documents/48598
17. Commission, E.: Proposal for a regulation of the European parliament and of the council on harmonised rules on fair access to and use of data (Data Act) (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0068
18. Commission, E.: Proposal for a regulation of the european parliament and of the council on the european health data space (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197
19. Crabtree, A., Mortier, R.: Human data interaction: historical lessons from social studies and CSCW, p. 20

20. Directive (EU) 2016/1148, 19.7.2016, O.L.: Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union. Official Journal of the European Union 194(59), 1–30 (2016). https://eur-lex.europa.eu/eli/dir/2016/1148/oj
21. Fette, G., Ertl, M., Störk, S.: Translating openEHR Models to FHIR. Studies in Health Technology and Informatics **270**, 1415–1416 (2020). https://doi.org/10.3233/shti200469
22. Frexia, F., et al.: openEHR is FAIR-Enabling by Design. preprint, Health Informatics (Feb 2021). https://doi.org/10.1101/2021.02.18.21251988, http://medrxiv.org/lookup/doi/10.1101/2021.02.18.21251988
23. Grönvall, E., Kyng, M.: On participatory design of home-based healthcare. Cognition, Technol. Work **15**(4), 389–401 (2013). https://doi.org/10.1007/s10111-012-0226-7
24. Grönvall, E., Lundberg, S.: Designing for offline and online social work: Technology-mediated collaborative practices in and between municipalities: European conference on cognitive ergonomics. European conference on cognitive ergonomics (2022). Association for Computing Machinery
25. Gøeg, K.R., Rasmussen, R.K., Jensen, L., Wollesen, C.M., Larsen, S., Pape-Haugaard, L.B.: A future-proof architecture for telemedicine using loose-coupled modules and HL7 FHIR. Computer Methods and Programs in Biomedicine **160**, 95–101 (2018). https://doi.org/10.1016/j.cmpb.2018.03.010, http://linkinghub.elsevier.com/retrieve/pii/S0169260717311707 13 Sept 2022. Elsevier B.V
26. International, O.: What is openEHR? (2021), shorturl.at/eoBOZ
27. Langley, J., Wolstenholme, D., Cooke, J.: 'Collective making' as knowledge mobilisation: the contribution of participatory design in the co-creation of knowledge in healthcare. BMC Health Services Research 18(1), 585 (2018). https://doi.org/10.1186/s12913-018-3397-y, https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-018-3397-y, tex.ids= langleyCollectiveMakingKnowledge2018a
28. Moharra, M., Almazán, C., Decool, M., Nilsson, A.L., Allegretti, N., Seven, M.: Implementation of a cross-border health service: physician and pharmacists' opinions from the epSOS project. Family Practice **32**(5), 564–567 (2015). https://doi.org/10.1093/fampra/cmv052. 13 Sept 2022
29. Mortier, R., Haddadi, H., Henderson, T., McAuley, D., Crowcroft, J.: Human-data interaction: the human face of the data-driven society (2015). http://arxiv.org/abs/1412.6159, arXiv:1412.6159
30. Pedrera-Jiménez, M., on EHR standards, S.E.G., Kalra, D., Beale, T., Muñoz-Carrero, A., Serrano-Balazote, P.: Can OpenEHR, ISO 13606 and HL7 FHIR work together? An agnostic perspective for the selection and application of EHR standards from Spain (2022). https://doi.org/10.36227/techrxiv.19746484.v1, 13 Sept 2022
31. Regulation (EU) 2016/679, O.L.: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Official Journal of the European Union 119(1), 1–88 (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj
32. Regulation (EU) 2022/868, O.L.: Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) (Text with EEA

relevance). Official Journal of the European Union 152(1), 1–44 (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020PC0767
33. Samarasinghe, S., Lokuge, S., Snell, L.: Exploring tenets of data democratization (2022). https://doi.org/10.48550/arXiv.2206.12051, http://arxiv.org/abs/2206.12051, arXiv:2206.12051
34. Seidelin, C., Dittrich, Y., Grönvall, E.: Co-designing data experiments, p. 11 (2020). https://doi.org/10.1145/3419249.3420152
35. Seidelin, C., Dittrich, Y., Grönvall, E.: Foregrounding data in co-design - an exploration of how data may become an object of design. Int. J. Human-Comput. Stud. **143**, 18 (2020). https://doi.org/10.1016/j.ijhcs.2020.102505, https://www.sciencedirect.com/science/article/pii/S1071581920301075
36. Simonsen, J., Robertson, T. (eds.): Routledge International Handbook of Participatory Design. Routledge (2012). https://doi.org/10.4324/9780203108543, https://www.taylorfrancis.com/books/9781136266263
37. Slegers, K., Wilkinson, A., Hendriks, N.: Active collaboration in healthcare design: participatory design to develop a dementia care app. In: CHI '13 Extended Abstracts on Human Factors in Computing Systems, pp. 475–480. CHI EA '13, Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2468356.2468440, 13 Sept 2022
38. Tcheng, J.E., et al.: Achieving Data Liquidity: Lessons Learned from Analysis of 38 Clinical Registries (The Duke-Pew Data Interoperability Project. AMIA ... Annual Symposium proceedings. AMIA Symposium 2019, 864–873 (2019)
39. Till, S., et al.: Community-based co-design across geographic locations and cultures: methodological lessons from co-design workshops in South Africa. In: Participatory Design Conference 2022: vol. 1, pp. 120–132. ACM, Newcastle upon Tyne United Kingdom (2022). https://doi.org/10.1145/3536169.3537786, https://dl.acm.org/doi/10.1145/3536169.3537786, 10 Sept 2022
40. Wang, Y., Blobel, B., Yang, B.: Reinforcing health data sharing through data democratization. J. Personalized Med. **12**, 1380 (2022). https://doi.org/10.3390/jpm12091380. 01 Sept 2022
41. Wilke, G., Portmann, E.: Granular computing as a basis of human–data interaction: a cognitive cities use case. Granular Comput. **1**(3), 181–197 (2016). https://doi.org/10.1007/s41066-016-0015-4

Facial Recognition, Gesture Recognition and Object Detection

A Quantitative Comparison of Manual vs. Automated Facial Coding Using Real Life Observations of Fathers

Romana Burgess[1,2](✉), Iryna Culpin[2,3], Helen Bould[2,4,5], Rebecca Pearson[2,3], and Ian Nabney[1]

[1] Digital Health Engineering Group, Faculty of Engineering, Merchant Venturers Building, University of Bristol, Bristol, UK
`romana.burgess@bristol.ac.uk`
[2] Centre for Academic Mental Health, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
[3] Department of Psychology, Manchester Metropolitan University, Manchester, UK
[4] Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK
[5] Gloucestershire Health and Care NHS Foundation Trust, Gloucester, UK

Abstract. This work explores the application of an automated facial recognition software "FaceReader" [1] to videos of fathers ($n = 36$), taken using headcams worn by their infants during interactions in the home. We evaluate the use of FaceReader as an alternative method to manual coding – which is both time and labour intensive – and advance understanding of the usability of this software in naturalistic interactions. Using video data taken from the Avon Longitudinal Study of Parents and Children (ALSPAC), we first manually coded fathers' facial expressions according to an existing coding scheme, and then processed the videos using FaceReader. We used contingency tables and multivariate logistic regression models to compare the manual and automated outputs. Our results indicated low levels of facial recognition by FaceReader in naturalistic interactions (approximately 25.17% compared to manual coding), and we discuss potential causes for this (e.g., problems with lighting, the headcams themselves, and speed of infant movement). However, our logistic regression models showed that when the face was found, FaceReader predicted manually coded expressions with a mean accuracy of $M = 0.84$ (*range* = 0.67–0.94), sensitivity of $M = 0.64$ (*range* = 0.27–0.97), and specificity of $M = 0.81$ (*range* = 0.51–0.97).

Keywords: Automated facial coding · FaceReader · ALSPAC

1 Introduction

Manual coding is a comprehensive method for capturing facial expressions from observational data, and is easily adaptable to a range of contexts and scenarios. However, manual coding is both time and labour intensive, and can be biased by the experiences of the human coder (for example, the amount of time spent coding, or the previous

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 379–396, 2023.
https://doi.org/10.1007/978-3-031-34586-9_25

expression coded). These disadvantages may be addressed by automated facial coding, which offers rapid and detailed decomposition of facial expressions. Automated facial coding could potentially cut down on time spent manually coding, and also reduce any biases. This would allow for more data to be processed, potentially more accurately.

One such automated facial coding software is the Noldus FaceReader [1] which was first proposed by Den Uyl & Van Kuilenburg [2], who described the process involved in finding, modelling, and classifying a face. The software uses deep learning and neural networks to classify faces into one of eight facial expressions: Happy, Sad, Angry, Scared, Surprised, Disgusted, Neutral and Contempt.

A validation study was performed on the software [3], comparing human and FaceReader facial classification for two publicly available, objective datasets of human expressions. On average, FaceReader correctly identified 89% of expressions, whereas human coders correctly identified 85%. This work also identified variation in accuracy across different expressions, e.g., FaceReader classified Happy with 96% accuracy, but classified Angry with 76% accuracy. Another validation was performed on a later version of the software [4], finding that - on average across all expressions - 80% of expressions were correctly classified. Other reported rates of performance for the software are 89% [2] and 87% [5]. One study observed gender-based differences [6], noting that FaceReader better identified Surprised and Scared emotions in males, and better identified Disgusted and Sad emotions in females.

Previous works have compared the performance of FaceReader to human coding. For example, one study investigated the expressions of students taking a test, measuring the agreement between two human coders and FaceReader [6]. This work found that the humans and the software agreed strongly for Neutral and Happy expressions, agreed often for Sad, Scared, and Surprised, and did not agree often for Disgust and Angry. A similar finding was reflected in [5], who found that agreement was highest between FaceReader and manual coders for Neutral and Happy, and lowest for Angry and Disgust. There have been many similar applications of the FaceReader software across different domains, including: evaluating human reactions to complex web-based tasks [7, 8], measuring implicit and explicit expressions during orange juice tasting sessions [9], and evaluating spontaneous expressions vs. posed expressions [10].

However, previous applications of FaceReader have commonly used videos filmed in controlled environments with good lighting, and a homogenous background with no people or objects [9–11]. To be useful for many real-life applications, it is vital that expression recognition is effective for naturalistic, uncontrolled environments, such as within the home. While some have implemented other methods to recognise faces for "in-the-wild" videos [12], we have found very few studies where FaceReader has been applied to these kind of observations. We identified one example [13] which used FaceReader to analyse a dataset containing movie clips of actors, and found that the software could not accurately classify any expression.

Many studies have also used videos that were recorded using built-in laptop webcams, providing a direct view of the participant's face. Naturalistic interactions – for example involving multiple people or different body positions – may not be well captured by a webcam, or any kind of stationary camera. Wearable headcams provide an ideal solution for capturing facial expressions in a naturalistic setting, and may enable more ecologically valid interactions, e.g., by containing less socially desirable facial expressions than in a controlled observation [14]. We have not identified any studies applying FaceReader analysis, or any other automated facial coding, to videos taken using wearable headcams during natural interactions.

FaceReader has rarely been explored in a parent-infant context. One study used the software to analyse the intensity of mothers' Happy expressions during exposure to images of infants [15], and another carried out five separate tests using FaceReader in a parent-infant context [16]. These tests analysed facial expressions across different scenarios, including mother-infant interactions, infant-infant interactions, and interactions in infants with developmental disorders. All except one test used observations carried out in naturalistic settings with uncontrolled lighting and a handheld video camera (one test used a laboratory setting with controlled lighting). Videos were excluded from analysis if the participants head rotation was greater than 45 degrees from the camera. The authors highlighted that assessing facial expressions in infant interactions is vital to understand the "emotional sphere" of the child, with the goal of identifying and addressing less optimal emotional responses (e.g., smiling at an infant cry) [16]. While these studies both contributed to understanding maternal expressions, we did not find any FaceReader literature specifically studying fathers' facial expressions.

Father-infant interactions have been studied much less than those between mothers and infants. This may be in part because fathers have traditionally been less involved in childcare, although modern fatherhood roles are evolving to include more social, emotional, and physical childcare than ever before [17]. Yet, studies of fathers are valuable, as father-infant interactions have been found to contribute to infant language and cognitive development [18] and to be predictive of behavioural problems [19]. Infants begin to develop emotional coordination – learning to discriminate and respond to emotional expressions, vital for social function [20] – from as early as 4 months old [21]. Whilst much is known about how emotional coordination occurs in mother-infant interactions, less is understood about fathers [22]. There are likely to be differences in the communicative mechanisms, and by observing, quantifying, and analysing father and infant facial expressions during an interaction, we can begin to understand these differences.

The aims of our work were: (1) to evaluate the performance of FaceReader on videos of naturalistic father-infant interactions captured using a wearable headcam, and (2) to evaluate the relationship between the automated and human coding of facial expressions. To address these aims, we coded 36 videos of fathers engaging in free play or feeding interactions with infants, both manually and using FaceReader. We then used contingency analysis and logistic regression classification models to compare the two relative outputs. Through this work, we provide new information regarding the use of FaceReader to process paternal facial expressions in a naturalistic setting.

2 Methodology

2.1 Data

We used data taken from the Avon Longitudinal Study of Parents and Children (ALSPAC). The study website (http://www.bristol.ac.uk/alspac/researchers/our-data/) contains details of all ALSPAC data that are available through a fully searchable data dictionary and variable search tool. ALSPAC data are collected and managed using Research Electronic Data Capture (REDCap) electronic data capture tools hosted at the University of Bristol [23]; REDCap is a secure web-based platform designed to support data capture for research studies.

Full ALSPAC cohort demographics and recruitment details have been provided previously elsewhere [24–27]. In brief, ALSPAC is an ongoing longitudinal, population-based study based in Bristol, UK. The original cohort was recruited via 14 541 pregnancies with expected delivery dates between 1 April 1991 and 31 December 1992; this original cohort is referred to as generation 0, or ALSPAC-G0. The children born in 1992 to the ALSPAC-G0 cohort are referred to as generation 1, or ALSPAC-G1. And finally, the children born to the ALSPAC-G1 cohort in recent years are referred to as generation 2, or ALSPAC-G2. Our work comprises videos of fathers from ALSPAC-G1 (whose infants are in ALSPAC-G2). The fathers in this work had a mean age of 31.31 years (*SD* = 5.45), and their infants had a mean age of 32.62 weeks (*SD* = 5.85). Eight infants were male, and five infants were female.

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time.

The videos were taken using headcams worn by infants during father-infant interactions within the home. Previous work has found first-person headcams to be reliable for capturing behaviours during parent-infant interactions [14]. Fathers were recruited through a father-specific research clinic, inviting dads to take part in several assessments when their child turned six months old. One of these assessments was the headcam study, for which there were no specific selection criteria. Table 1 outlines the video contributions from each father included in our work.

Table 1. Detail about the videos provided by each father within the dataset.

ID	# Videos	Total length (sec)	Interaction types (# of videos)
F1	1	789	Feeding (1)
F2	2	736	Feeding (1), Free play (1)
F3	2	1141	Feeding (1), Free play (1)
F4	6	2776	Feeding (4), Free play (2)
F5	3	1656	Feeding (3)
F6	1	430	Free play (1)
F7	6	2325	Feeding (3), Free play (2), Combination (1)
F8	4	1396	Feeding (2), Free play (1), Combination (1)
F9	2	782	Feeding (1), Free play (1)
F10	1	916	Feeding (1)

(*continued*)

Table 1. (*continued*)

ID	# Videos	Total length (sec)	Interaction types (# of videos)
F11	3	2136	Feeding (2), Free play (1)
F12	2	1263	Feeding (2)
F13	3	1022	Feeding (1), Free play (1)
Total	**36**	**17 368**	**Feeding (24), Free play (10), Combination (2)**

We used 36 videos in total, collected between 2019–2020, including: 24 feeding interactions, 10 free play interactions, and 2 videos which were a combination of both feeding and free play. These videos come from 13 individual fathers, as many provided multiple separate videos (see Table 1).

As this work is not about individual differences in expressions, but rather the efficacy and accuracy of the software at capturing expressions, we found no reason to exclude second or third videos from a single participant. Also, as all videos were varied in length, it was possible that one father may include multiple videos roughly equating to the length of a single video from another father (e.g., F1 and F2).

2.2 Manual Coding

All videos were manually coded using Noldus Observer 15 software [28], at a temporal resolution of 1/5 s. The facial expressions were coded according to the MHINT coding scheme [29], which is freely available to access online. Specifically, this includes the following expressions: Smile, Positive, Neutral/Alert, Negative, Surprise, Mock Surprise, Woe Face, Disgust, None of the Above and Face not Visible. These expressions are exhaustive and mutually exclusive, meaning that every timestamp is allocated a unique, single expression.

Initial coding was performed by one researcher, and two additional researchers were recruited for double coding. Seven randomly selected videos were selected for double coding, with one researcher coding four videos, and one researcher coding three. This equated to 3906 s of double-coded data, or 22.38% of the total video data. All reliability analyses were conducted using the Observer XT 15.0 [28].

To measure inter-coder agreement, we used the *index of concordance*. This is calculated by the total agreement for a behaviour (i.e., the duration that an expression is coded as present/not present by both coders) divided by the total duration of the interaction. The index of concordance is expressed as a value between 0 (no agreement) and 1 (total agreement). Across all expressions, an index of concordance of 0.93 ($SD = 0.07$) was achieved with the first double coder, and 0.91 ($SD = 0.07$) was achieved with the second coder. Inter-coder results by facial expression are shown in Table 2. In these analyses, we excluded expressions that occurred for less than 1% of the total interaction duration (i.e., Woe face, Disgust, Surprise).

Table 2. Mean inter-coder reliability by facial expression (*SD*).

Neutral	Positive	Smile	Negative	Mock Surprise	None of the above	Face not visible
0.89 (0.09)	0.89 (0.09)	0.96 (0.03)	0.98 (0.00)	0.98 (0.02)	0.94 (0.05)	0.89 (0.08)

2.3 Automated Facial Coding Using FaceReader

All videos were also processed using Noldus FaceReader [1], a facial recognition software trained to classify eight facial expressions in adult humans: Happy, Sad, Neutral, Angry, Scared, Surprised, Disgusted, Contempt. To make this classification, deep learning algorithms are first used to find a face within an image, while varying for facial position and size. Eye tracking is also used to help identify the rotation of the face. Based on deep neural networks, an artificial face is then synthesised using around 500 key points on the face; these describe the position of features and muscles, as well as different textures (e.g., eyebrow presence).

Expression classification takes place using a neural network, trained to recognise facial expressions using over 20,000 manually annotated images of faces [30]. When presented with a face, the software fits a "mesh" over the face and its key points, then calculates the deviation of these points from their position relative to a "mean" face, in order to make a prediction of the expression. A more detailed explanation of the FaceReader software can be found elsewhere [31].

Once a face has been detected, FaceReader provides multiple detailed outputs describing the facial expression present within that frame. The software processes videos frame-by-frame, which in our case resulted in an output being provided for every 0.033 s. In this work, we use the FaceReader output *expression intensity*: a single value within the interval [0, 1] describing the strength of each of the eight expressions. An intensity close to 0 indicates the expression is not present, and an intensity close to 1 indicates an expression is very present. An intensity value is provided for each of the eight expressions simultaneously, with each value independent of one another (i.e., the values do not sum to one).

It should be noted that the FaceReader expressions do not directly match those within the manual coding scheme (e.g., "Mock Surprise" is a manual expression, but not a FaceReader one), however, this was not problematic for the purposes of our work. Table 3 shows an approximate mapping between the manual and FaceReader expressions.

Table 3. Approximate mapping between manual and FaceReader expressions.

Manual Expression(s)		FaceReader Expression(s)
Neutral/Alert	→	Neutral
Smile + Positive	→	Happy
Negative	→	Sad + Angry + Scared + Contempt
Surprise + Mock surprise	→	Surprised
Disgust	→	Disgusted
None of the Above + Woe face	→	*n/a*
Face not visible	→	Face not found

2.4 Data Analysis

Data Pre-processing. Before we started the data analysis, we first carried out some pre-processing. A flow diagram showing all pre-processing stages is provided in the Appendix. We started by removing all data where a second caregiver was present during the interaction. This typically happened when the mother came to bring food, to admire the headcam on the infant, or to walk past in the background. These data were removed because FaceReader would often mistakenly classify the facial expression of the second caregiver during these periods, rather than that of the father. This meant that amount of viable coded data reduced from 17,368 s to 15,420 s.

The expression intensities were also normalised during pre-processing. This was necessary because the manual coder can only choose one dominant facial expression at a time, so for consistency, we must assume that we cannot have multiple dominant expressions. By normalising the intensities, this helps to highlight the dominant expression within the FaceReader output.

Data Analysis Procedures. All analyses were carried out using Python 3.0 [32]. Our aims were twofold: (1) to evaluate the performance of FaceReader on videos of naturalistic father-infant interactions captured using a wearable headcam, and (2) to evaluate the relationship between the automated and human coding of facial expressions.

To address aim (1), we calculated the amount of time that a face was detected and classified by both the human coder and the FaceReader software. We also calculated the amount of time that a face was not detected by both the human and the software. These values are displayed in a contingency table (Table 4) in Sect. 3.1.

To address aim (2), we used multivariate binary logistic regression; a choice made due to the simplicity of the model and the ease of parameter interpretation. Logistic regression measures the probability of a data entry being classified as one of two mutually exclusive, exhaustive states (which we assign as either a 0 or a 1). In our work, this translates to the eight simultaneous, FaceReader expression intensities being classified as one of the manually coded facial expressions (classified as a 1) or not (classified as a 0). Employing multivariate binary logistic regression meant that a separate logistic regression model was implemented for each of the manually coded facial expressions. The process for fitting a single model is outlined below:

1. Split the dataset into a train and test set. Here, all data for a single person must be contained in either the train or the test set (a father cannot be within both). This helps to avoid inflated prediction measures of generalization performance. We used 13 fathers in total: 10 of these comprised the training set ($n = 44{,}352$ frames), and 3 fathers comprised the test set ($n = 5{,}180$ frames).
2. Define our features $\boldsymbol{X}$ (the normalised FaceReader expression intensities) and our target variable $\boldsymbol{Y}$ (the manually coded facial expression). $\boldsymbol{X}$ is an $8 \times n$ matrix, where 8 is the number of FaceReader expressions, and n is the number of entries in the training dataset. $\boldsymbol{Y}$ is a binary array (1 indicates the manual facial expression of interest, 0 indicates any other expression) of length n, where n is the number of entries in the training dataset.
3. Using the Python package sk-learn, fit the logistic regression model on the training data. We used the LBFGS solver (an optimisation algorithm approximating the Broyden-Fletcher-Goldfarb-Shanno algorithm, see [33]), and we weighted the classes based on their frequencies in the training data, adjusting for the imbalance of behaviours per class (see Fig. 1).
4. Test the fitted model using the test set.

The fathers in the test and train datasets were selected through a trial-and-error process, with the aim of retaining a similar percentage of data per expression in each dataset, subject to the requirement of having all data from a single participant in only one dataset. The resulting representation of each manually coded expression in the full dataset, the training set, and the testing set is shown in Fig. 1 below.

We produced similar representation across almost all retained expressions (except for Mock Surprise, which was more heavily weighted in the test set). Following this evaluation, we excluded prediction models for Disgust, Surprise and Woe face, as these expressions each accounted for $< 1\%$ of the data.

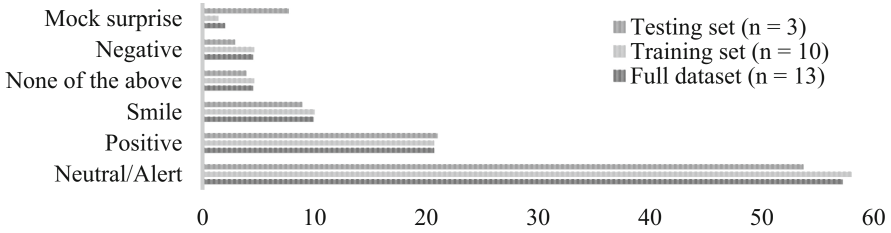


Fig. 1. Percentage of class occurrence in each dataset (%). Surprise, Woe face and Disgust were removed due to low prevalence ($< 1\%$).

3 Results

3.1 Quantifying FaceReader Performance Compared to Manual Coding

To address aim (1), we calculated how frequently the FaceReader software found the participant's face compared to the human coder. Our findings are provided in the contingency table below. Throughout our discussion, we assume the manually coded expressions to be correct, such that the manual expression is used as a benchmark with which the FaceReader output is compared.

Table 4. Face found vs. not found by the FaceReader software and the manual coder (%).

Manual	FaceReader (%)	
	Face Found	Face not found
Face found	10.71	31.84
Face not found	0.47	57.00

We found that throughout the 15,420 s of data, the manual coder identified a facial expression 42.55% of the time (from the table, manual face found = 10.71 + 31.84), while FaceReader only identified a facial expression 11.18% of the time (from the table, FaceReader face found = 10.71 + 0.47). Additionally, when a face was actually present (and was therefore coded by the human), FaceReader found the face 25.17% of the time (calculated by face found by both / total manual face found = 10.71 / (10.71 + 31.84) = 10.71 / 42.55). Table 4 also shows that percentage of time where FaceReader found a face but the Manual coder did not is very low (0.47%).

The observations where both FaceReader and the human coder classified a facial expression comprise the datasets used for the logistic regression analyses, i.e., we did not include data where the face was not found by either FaceReader or the human coder. The observations coded by both the human and the software comprise 1651 s of data, or 49,532 frames approximately.

3.2 The Relationship Between the Automated and Human Coding of Facial Expressions

To address aim (2), we looked to understand how accurately FaceReader facial expression intensities predicted manually coded facial expressions. The methodology for these analyses has been outlined in Sect. 2.4.

We fit six binary logistic regression models (one for each expression) to the training set, before testing the models on the test set. The resulting accuracy, sensitivity and specificity measures are provided in Table 5.

Accuracy represents the proportion of correct predictions for both classes; high accuracy means that the number of correct predictions (of either a 1 – the expression is present – or a 0 – the expression is not present) is high. Our six models all showed accuracy greater than 0.67, with most being greater than 0.80. The most accurate models were

Table 5. Accuracy, Sensitivity and Specificity for the logistic regression models.

Class	Predictive Performance Measure		
	Accuracy	Sensitivity	Specificity
Neutral ($n = 25{,}713$)	0.78	0.97	0.51
Positive ($n = 9172$)	0.67	0.48	0.72
Negative ($n = 2099$)	0.82	0.83	0.82
Smile ($n = 4437$)	0.94	0.84	0.94
None of the above ($n = 2021$)	0.88	0.27	0.91
Mock surprise ($n = 606$)	0.94	0.47	0.97

Mock Surprise and Smile (0.94), followed by None of the Above (0.88) and Negative (0.82). The prediction model for Positive (0.67) was the least accurate, followed by Neutral/Alert (0.78).

Sensitivity represents the model's ability to predict a true positive (i.e., correctly classifies a facial expression as present); high sensitivity means that the model rarely incorrectly classifies an expression as not present, and performs well at correctly classifying a facial expression as present). The sensitivity for some models were very high, including for Neutral (0.97), Smile (0.84) and Negative (0.83). However, the other models performed poorly: sensitivity for None of the Above was very low (0.27), with Mock Surprise (0.47) and Positive (0.48) only performing slightly higher.

Specificity represents the model's ability to predict a true negative (i.e., correctly class a facial expression as not present); high specificity means that the model rarely incorrectly classifies an expression as present), and performs well at correctly classifying facial expressions as not present). Our specificities were nearly all greater than 0.80. The highest specificity was for Mock Surprise (0.97), followed by Smile (0.94), None of the Above (0.91), Positive (0.82) and Negative (0.82). The lowest specificity was for Neutral/Alert (0.51).

4 Discussion

4.1 Summary of Results

Our work used 36 videos of fathers taken using headcams worn by infants during parent-infant interactions, totaling 15,420 s of data (after pre-processing). We manually coded the videos according to an established facial expression coding scheme [29], and also using an automated facial coding software. We applied contingency analysis to calculate how frequently FaceReader detected a face at the same time the human coder did, and used multivariate logistic regression models to evaluate the relationship between automated facial classification and manual coding.

Our results showed that FaceReader only found a facial expression around a quarter of the time that a human coder did (25.17%). This is not surprising, as it has been established that automated facial recognition is disadvantaged in real-world conditions

[13]. As such, we regard this low FaceReader performance as indicative of the high ecological validity of the interactions. Whilst previous studies have excluded data that FaceReader had failed to analyse, for reasons of signal loss during facial recognition [34], it was important that we retained these data in order to validate the software for naturalistic observations.

Our logistic regression models found that FaceReader predicted the manually coded expressions with generally good accuracy (*mean* = 0.84, *range* = 0.67–0.94), sensitivity (*mean* = 0.64, *range* = 0.27–0.97), and specificity (*mean* = 0.81, *range* = 0.51–0.97). The high accuracies across the six models suggest that our models were good at predicting when expressions were present. FaceReader was most accurate for Mock Surprise (0.94) and Smile (0.94). This is similar to [11], who found that FaceReader most accurately detected Surprise compared to other expressions. We found that FaceReader was least accurate for Positive (0.67) and Neutral/Alert (0.78), a finding inconsistent with other studies [5, 11, 13]. This is particularly interesting in the case of [13], who trialed FaceReader for in-the-wild facial detection, similarly to us.

We found high sensitivities for the Neutral/Alert, Smile and Negative models, meaning that FaceReader was able to positively identify the presence of these expressions. Conversely, FaceReader was less accurate in identifying None of the Above or Positive. This could be caused by a high number of false negative predictions, i.e., failing to predict an expression that was in fact present. This makes sense in the case of None of the Above, which serves to represent an “other” category, meaning that it encompassed multiple different expressions. In our work, this was mostly fathers eating, sneezing, yawning, or opening their mouths to mimic eating food, but could have included an even wider range of expressions. As such, we would not expect sensitivity to be as high for None of the Above as for the other expressions. For Mock Surprise and Positive, however, low sensitivity indicates that for some reason, our models often failed to correctly identify these expressions.

We generally found very high specificities (all were greater than 0.80 except for Neutral/Alert), meaning that FaceReader performed well at correctly identifying that a particular expression was absent. The low specificity for Neutral/Alert (0.51) indicates that the false positive rate was high – i.e., FaceReader incorrectly identified many expressions as Neutral. This is similar to [35], who reported that their low accuracy for Neutral (19%) was due to FaceReader over-detecting neutral expressions.

4.2 Failures in Face Detection

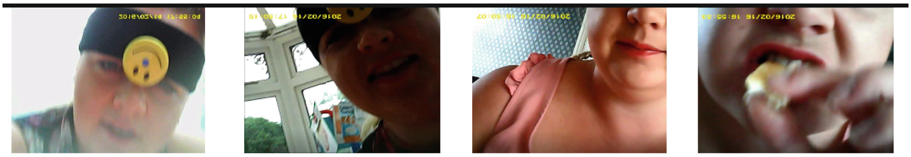
Previous work has highlighted the importance of creating more naturalistic settings for FaceReader applications [36]. In practice, naturalistic settings are problematic for a successful FaceReader analysis. In our work, the software struggled to detect a face across many scenarios where the human coder had no trouble. We looked to our videos to identify the reasons for this, and provide some examples here. Figure 2 shows some specific reasons that we believe FaceReader performance was low in our videos, including: blurry images (a), bad lighting (b), the face being partially out of shot (c and h), headcams/toys/food blocking the face (d and f), the parent facing away from the camera (e), or FaceReader misclassifying another object as a face (g). For confidentiality reasons, we are not able to share images of the fathers used within this study. However, the images in Fig. 2 show a mother from a mother-specific, but otherwise identical, headcam study (this mother consented to have her data shared).

FaceReader documentation [31] outlines that the software performs less well if: the participant is wearing glasses, the lighting in the room is too dark, the participant is rotated away from the camera, or something is partially blocking the face (e.g., thick facial hair, hands, or a hat). Issues with glasses, bodily occlusions and artificial illumination have also been reported by others [13, 37]. The videos in our work were collected for an earlier study, so we were not able to advise that participants avoided wearing hats or glasses, that they sat in a room with natural lighting, or that they restrained from blocking their face with objects. It is therefore unsurprising that we observed a lot of these issues in the videos. Of our 13 fathers, six had facial hair of varying thickness, and three wore glasses. Additionally, many interactions also took place in front of a window, in rooms that were artificially lit, or in dim early morning or evening light, which lead to problems with lighting on fathers' faces (see Fig. 2b). It is possible that these factors lessened FaceReaders ability to analyse the faces in our videos.

While there are many reasons why FaceReader struggles to locate the face (Fig. 2), we suggest that the types of interactions we studied may also have affected this. In the free play interactions ($n = 10$), there were fast movements and variation between different positions (e.g., sit, lie on front), which led to images being both sporadic and blurry (see Fig. 2a). This also meant that the infant was not always at eye level with the father, meaning that the top, bottom, or side of the face was often out of shot (see Fig. 2c). Further, even if the full face was in sight, sometimes toys used for play would block the view of the face. While a human coder may be able to distinguish a facial expression in spite of small obstructions, slightly blurry shots, and missing parts of the face, FaceReader would not.

Similarly, the context of a Feeding interaction ($n = 24$) meant that cutlery, bowls, or food were often raised in front of the infant's face, partially or wholly blocking view of the father (see Fig. 2d). Further, while feeding interactions generally involved fathers sitting facing their infant in order to feed them, there were instances where the father was also eating a meal. This led to sideways or otherwise indirect views of the father (see Fig. 2e), if he was sitting next to the infant or somewhere around the table. While a human coder may distinguish a facial expression from a sideways or angled view, FaceReader is not able to do this beyond around a 40-degree tilt [31].

We also suggest that the use of the headcam led to a reduced FaceReader performance. For example, one problem we encountered was that the headcam was placed too high on the infant's head (i.e., pointing more upwards than forwards), so that only the top of the father's head was visible. Conversely, a headcam that was placed too low on the father's face meant that his eyebrows were covered, causing the face to be uncaptured by FaceReader (see Fig. 2f). In both of these cases, the face was typically still visible to the human coder. In a previous study investigating headcam use for capturing dyadic interactions [38], the experimenters manually adjusted the subject headcams to ensure that they were well fitted, and positioned to capture the most desirable perspective. In our work, subjects put the headcams on themselves, which meant that there was more likely to be placement issues.

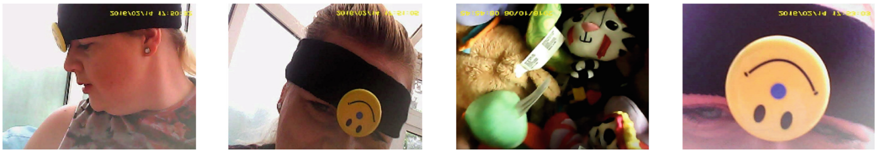


a. The image is blurry. This may happen if the infant (or father) moves their head too quickly.

b. The lighting in the room is bad, e.g., the parent is sitting in front of a window, or there is a lack of natural light.

c. Any part of the face is missing from the image. This may happen if the parent moves from the infant's line of sight, or the infant looks elsewhere.

d. A hand, toy or food-related item is blocking the view of the parent's face.



e. The parent is facing sideways, upwards, or downwards. This can obstruct facial features from view.

f. The headcam is blocking the view of the face, i.e., by being worn too low on the head, and covering the eyebrows.

g. A face that is not the parent's face is recognised and classified, (e.g., a face on a t-shirt, a stuffed animal, a painting, or another person.)

h. The parent is too close. This may happen when the parent leans in (e.g., to feed or kiss the infant). The face may also be unclassified if the parent is too far away.

Fig. 2. Examples of images where FaceReader does not find the face. Images are taken from an infant headcam during mother-infant interactions in a near identical study.

Finally, FaceReader also provided outputs for some frames that the human coder did not (0.47%). There are many reasons that this could happen, including where FaceReader misclassifies: a different person's face, an item of clothing or poster with a face on, or some other background object (see Fig. 2g). Complex backgrounds have previously been reported as problematic [13]. To account for this effect, we had already removed data from our analyses that included another caregiver in the background of the video, meaning that these misclassifications were likely to be caused by clothing or background items.

4.3 Implications for Future Work

To our knowledge, no previous studies have used FaceReader to analyse headcam videos. It would therefore be useful to further evaluate how headcams can be best used to optimise the chances of a successful FaceReader analysis. This is particularly important as the use of headcams allows for a more ecologically valid and natural observation [14]. We therefore outline six recommendations for future work in this area, aiming to maintain the authenticity of a natural observation, while also optimising the logistical aspects of the observation.

[1] Subjects should be instructed on how to properly put the headcams on both themselves and the infant (to avoid pointing up/down and missing the partners face).
[2] Subjects should be advised regarding optimal lighting conditions (i.e., natural light). Where possible, observations should take place in an area of natural lighting, preferably during the middle of the day where natural light is most prevalent.
[3] Subjects should be advised regarding glasses (not to wear them if possible), or other facial occlusions. Whilst no hands in front of the face would be desirable, highlighting this could affect the authenticity of subject behaviours and movements, so we would not recommend mentioning this to subjects.
[4] Depending on the interaction, researchers could advise subject posture (e.g., we found that feeding interactions had successful analysis when subject was face to face with infant). It would be beneficial to suggest interactions or observations that naturally cause subjects to face the camera front-on.
[5] Subjects should not sit in view of photos or posters hung on walls, and should not wear t-shirts or other clothing items with people shown on them.
[6] Researchers should support the development and usage of more powerful compact cameras, to provide greater robustness against rapid head movements (as previ-ously suggested by [9])

While these recommendations should aid in optimising FaceReader performance for naturalistic interactions, the reality is that – for now – it may be necessary that manual coding (or some other method) is used to supplement FaceReader coding. In our case, this would account for roughly 25% of faces being automatically coded, and the remaining having to be supplemented by a human coder (this is, if the end goal is 100% coding). However, it may be that this is sufficient for drawing useful conclusions in many cases, especially when using high quality, long extracts of video data. While this is not nearly close to the goal of fully automated facial coding, 25% represents a starting benchmark which future work can aim to improve upon.

Additionally, FaceReader provides much more detail than what is capable from a human coder, i.e., expression intensity for eight concurrent expressions. This means that even with a performance rate of 25%, we can potentially learn a lot more from the FaceReader output that wouldn't be possible from manual coding alone. In a clinical scenario, therefore, where manual coding is impractical, it is easy to see the potential utility of automated coding techniques, even at a 25% performance rate.

Future work could also identify whether the successful automated coding is biased towards certain expressions (e.g., expressions may be more prevalent when FaceReader is unsuccessful, such as when the second caregiver is present, or when the parent or infant turns toward a distraction). Similarly, it would be beneficial to investigate how FaceReader could work alongside complementary techniques (e.g., linear interpolation) to identify "missing" facial expressions.

4.4 Strengths and Limitations

A major strength of our research is the use of real-life observations, which meant that fathers exhibited natural, unposed facial expressions. Without the presence of a third-party researcher to record the interactions, it is likely that the recordings captured more ecologically valid facial expressions [14]. Also, using headcams during a dyadic interaction allowed the camera to focus directly on fathers' faces. This is an advantage over third person cameras, which can miss out on capturing facial expressions [14].

For limitations, we acknowledge there was low prevalence of some facial expressions, meaning that some had to be excluded from the logistic regression models (e.g., Woe face, Disgust). Additionally, some of the models possibly did not perform as well as they might have with more data. The facial expressions are a direct reflection of fathers' emotions while interaction with their children, however, so it is not surprising that we did not encounter lots of Disgust, for example.

Further, there was not a one-to-one relationship between the facial expressions in the manual coding scheme and those implemented by FaceReader. For example, the manual coding scheme contained the expression Negative, while FaceReader contained the separate negative expressions Sad, Scared, Angry and Contempt. If we had manually-coded these negative expressions separately, it is possible that we may have been able to implement better performing, distinct models for each expression. Having said that, the prevalence of Negative was quite low ($n = 2099$) compared to other expressions (e.g., Neutral/Alert, Smile), suggesting that there may not have been quite enough data to have created three, high performing, separate models in this instance.

As previously acknowledged, the length of video material for each subject was not the same (i.e., we had many more facial expressions for some fathers than others, especially where some fathers provided multiple videos). It is possible that this led to biases in how our models learned to predict certain facial expressions. However, we modulated for this effect in our performance measures by including all data from a given subject in either the training or the testing dataset.

There are some inherent biases present within the ALSPAC cohort, many of which have been detailed previously [26]. One example is that the cohort is mostly of White-European origin, reducing the generalisability of findings to the general population. However, as our work does not aim to interpret specific facial expressions present in the interactions, generalisability of findings is not as important here as for other studies.

Finally, we acknowledge that wearing headcams might have influenced the natural facial expressions of our participants. Although, we believe that the effect of this was mitigated (and countered) by the more ecologically valid expressions that we expected to see from real life, at-home interactions.

Acknowledgements. We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. We would also like to thank the two double coders who were involved in this work, Lottie Relph and Maddy Stephens.

This publication is the work of the authors RB, IC, HB, and RP, who will serve as guarantors for the contents of this paper. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf); This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 758813; MHINT).

Additionally, RB was supported by the Engineering and Physical Sciences Research Council (EPSRC) Digital Health and Care Centre for Doctoral Training (CDT) at the University of Bristol (UKRI Grant No. EP/S023704/1). IC was supported by the Wellcome Trust Research Fellowship in Humanities and Social Science (Grant ref: 212664/Z/18/Z).

Appendix

Here we provide a flow diagram to demonstrate how we processed the data involved in this study.

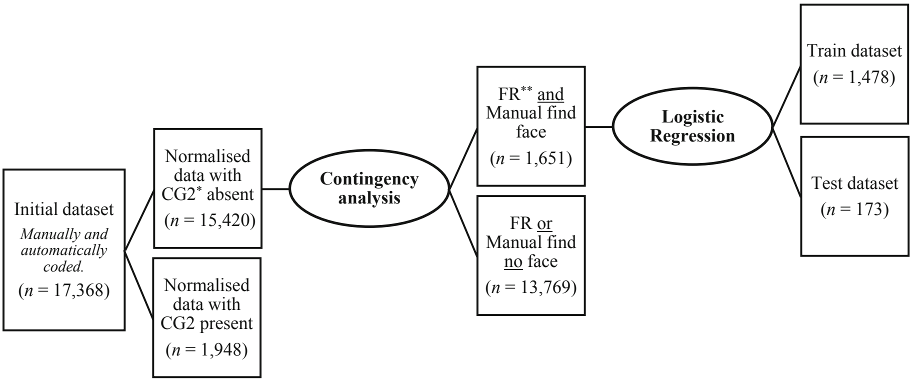


Fig. 3. Flow diagram to demonstrate stages of data processing; *n* refers to the number of video frames. *CG2 = Caregiver 2. **FR = FaceReader.

References

1. Noldus: FaceReader (2022). https://www.noldus.com/facereader
2. Den Uyl, M.J., Van Kuilenburg, H.: The FaceReader: online facial expression recognition. In Proceedings of Measuring Behavior, vol. 30, no. 2, pp. 589–590. Wageningen (2005)
3. Lewinski, P., den Uyl, T.M., Butler, C.: Automated facial coding: validation of basic emotions and FACS AUs in FaceReader. J. Neurosci. Psychol. Econ. **7**(4), 227 (2014)
4. Skiendziel, T., Rösch, A.G., Schultheiss, O.C.: Assessing the convergent validity between the automated emotion recognition software Noldus FaceReader 7 and facial action coding system scoring. PLoS ONE **14**(10), e0223905 (2019)
5. Terzis, V., Moridis, C.N., Economides, A.A.: Measuring instant emotions based on facial expressions during computer-based assessment. Pers. Ubiquit. Comput. **17**(1), 43–52 (2013)

6. Terzis, V., Moridis, C.N., Economides, A.A.: Measuring instant emotions during a self-assessment test: the use of FaceReader. In: Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research, pp. 1–4 (2010)
7. Talen, L., den Uyl, T.E.: Complex website tasks increase the expression anger measured with FaceReader online. Int. J. Human–Comput. Interact. 1–7 (2021)
8. Zaman, B., Shrimpton-Smith, T.: The FaceReader: measuring instant fun of use. In: Proceedings of the 4th Nordic conference on Human-Computer Interaction: Changing Roles, pp. 457–460 (2006)
9. Danner, L., Sidorkina, L., Joechl, M., Duerrschmid, K.: Make a face! Implicit and explicit measurement of facial expressions elicited by orange juices using face reading technology. Food Qual. Prefer. **32**, 167–172 (2014)
10. Bența, K.I., et al.: Evaluation of a system for realtime valence assessment of spontaneous facial expressions. In: Distributed Environments Adaptability, Semantics and Security Issues International Romanian-French Workshop, Cluj-Napoca, Romania , pp. 17–18 (2009)
11. Brodny, G., Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., Wróbel, M.R.: Comparison of selected off-the-shelf solutions for emotion recognition based on facial expressions. In: 2016 9th International Conference on Human System Interactions (HSI), pp. 397–404. IEEE (2016)
12. Krishna, T., Rai, A., Bansal, S., Khandelwal, S., Gupta, S., Goyal, D.: Emotion recognition using facial and audio features. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 557–564 (2013)
13. Gómez Jáuregui, D.A., Martin, J.C.: Evaluation of vision-based real-time measures for emotions discrimination under uncontrolled conditions. In: Proceedings of the 2013 on Emotion Recognition in the Wild Challenge and Workshop, pp. 17–22 (2013)
14. Lee, R., et al.: Through babies' eyes: practical and theoretical considerations of using wearable technology to measure parent–infant behaviour from the mothers' and infants' viewpoints. Infant. Behav. Dev. **47**, 62–71 (2017). https://doi.org/10.1016/j.infbeh.2017.02.006
15. Karreman, A., Riem, M.M.: Exposure to infant images enhances attention control in mothers. Cogn. Emot. **34**(5), 986–993 (2020)
16. Lyakso, E., Frolova, O., Matveev, Y.: Facial Expression: psychophysiologcal study. In: Handbook of Research on Deep Learning-Based Image Analysis Under Constrained and Unconstrained Environments, pp. 266–289. IGI Global (2021)
17. O'Brien, M.: Shared caring: bringing fathers into the frame (2005)
18. Tamis-LeMonda, C.S., Shannon, J.D., Cabrera, N.J., Lamb, M.E.: Fathers and mothers at play with their 2-and 3-year-olds: Contributions to language and cognitive development. Child Dev. **75**(6), 1806–1820 (2004)
19. Ramchandani, P.G., Domoney, J., Sethna, V., Psychogiou, L., Vlachos, H., Murray, L.: Do early father–infant interactions predict the onset of externalising behaviours in young children? Findings from a longitudinal cohort study. J. Child Psychol. Psychiatry **54**(1), 56–64 (2013)
20. Feldman, R.: Infant–mother and infant–father synchrony: the coregulation of positive arousal. Infant Mental Health J. Official Publ. World Assoc. Infant Mental Health **24**(1), 1–23 (2003)
21. Montague, D.P., Walker-Andrews, A.S.: Peekaboo: a new look at infants' perception of emotion expressions. Dev. Psychol. **37**(6), 826 (2001)
22. Kokkinaki, T., Vasdekis, V.G.S.: Comparing emotional coordination in early spontaneous mother–infant and father–infant interactions. Eur. J. Develop. Psychol. **12**(1), 69–84 (2015)
23. Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., Conde, J.G.: Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. **42**(2), 377–381 (2009). https://doi.org/10.1016/j.jbi.2008.08.010

24. Boyd, A., et al.: Cohort profile: the 'children of the 90s'; the index offspring of the avon longitudinal study of parents and children (ALSPAC). Int. J. Epidemiol. **42**, 111–127 (2013). https://doi.org/10.1093/ije/dys064
25. Fraser, A., et al.: Cohort profile: the avon longitudinal study of parents and children: ALSPAC mothers cohort. Int. J. Epidemiol. **42**, 97–110 (2013). https://doi.org/10.1093/ije/dys066
26. Lawlor, D.A., et al.: The second generation of the Avon longitudinal study of parents and children (ALSPAC-G2): a cohort profile. Wellcome open research, 4, 36 (2019). https://doi.org/10.12688/wellcomeopenres.15087.2
27. Northstone, K, et al.: The Avon longitudinal study of parents and children (ALSPAC): an update on the enrolled sample of index children in 2019. Wellcome Open research, 4:51 (2019). https://doi.org/10.12688/wellcomeopenres.15132.1
28. Noldus. The Observer XT (2022a). http://www.noldus.com/human-behavior-research/products/the-observer-xt
29. Costantini, I., et al.: Mental health intergenerational transmission (MHINT) process manual (2021). https://doi.org/10.31219/osf.io/s6n4h
30. Gudi, A.; Tasli, H.E.; Den Uyl, T.M.; Maroulis, A.: Deep learning based facs action unit occurrence and intensity estimation. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) 4 May 2015, vol. 6, pp. 1–5. (2015)
31. Loijens, L., Krips, O., Grieco, F., van Kuilenburg, H., den Uyl, M., Ivan, P.: FaceReader 8 reference manual, noldus information technology (2020)
32. Van Rossum, G., Drake, F.L.: Python 3 reference manual. scotts valley, CA: CreateSpace (2009)
33. Fletcher, R.: Practical Methods of Optimization. John Wiley & Sons, Hoboken (2013)
34. Weth, K., Raab, M.H., Carbon, C.C.: Investigating emotional responses to self-selected sad music via self-report and automated facial analysis. Music. Sci. **19**(4), 412–432 (2015)
35. Matlovic, T., Gaspar, P., Moro, R., Simko, J., Bielikova, M.: Emotions detection using facial expressions recognition and EEG. In: 2016 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), pp. 18–23. IEEE (2016)
36. Booijink, L.I.: Recognition of emotion in facial expressions: the comparison of FaceReader to fEMG and self-report (Master's thesis) (2017)
37. Webber, M.: Can jealousy be detected as a unique pattern of recordable facial expressions by the FaceReader, and thus do such expressions manifest differently between sexes upon exposure to jealousy–evoking Snapchat messages?" (2018)
38. Park, C.Y., et al.: K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. Sci. Data **7**(1), 1–16 (2020)

Knowledge-Driven Dialogue and Visual Perception for Smart Orofacial Rehabilitation

Jacobo López-Fernández(✉), Luis Unzueta, Meritxell Garcia, Maia Aguirre, Ariane Méndez, and Arantza del Pozo

Fundación Vicomtech, Basque Research and Technology Alliance (BRTA),
Mikeletegi 57, 20009 Donostia-San Sebastián, Spain
{jlopez,lunzueta,mgarciap,magirre,amendez,adelpozo}@vicomtech.org

Abstract. This paper addresses the problem of accomplishing Orofacial Rehabilitation (OR) with the assistance of artificial intelligence. The main challenges involve accurately monitoring and interacting with the trainees, while preserving user experience. We analyse different approaches to solving these challenges and propose a methodology to build smart knowledge-driven OR systems that focus on automated interaction. Our proposal leverages the combination of vision-based micro and macro facial expression recognition and skill-based dialogue systems, which facilitate encapsulating the knowledge of rehabilitation professionals into natural language interactions. Experimental results of spoken keyword spotting and micro and macro facial expression recognition algorithms are provided. The OR expressions image dataset employed in our experiments is also published to support further research in the field.

Keywords: Orofacial Rehabilitation · Dialogue Systems · Facial Expression Recognition

1 Introduction

Within an ageing society it is common for individuals to develop physical or cognitive detriment. These impairments typically impact on life quality and people frequently need to be provided with rehabilitation services. Orofacial Rehabilitation (OR) is the branch of Physiatry dedicated to mitigating physical impairments in the orofacial system, which is the set of organs responsible for the physiological functions of breathing, sucking, swallowing, speaking and phonation, including all kinds of facial expressions [21]. Examples of facial gestures for OR are: *bite lower lip*, *bite upper lip*, *blink*, *blow cheeks*, *blow left cheek*,

---

Supported by SHAPES – Smart and Health Ageing through People Engaging in Supportive Systems - is funded by the Horizon 2020 Framework Programme of the European Union for Research Innovation. Grant agreement number: 857159 - SHAPES – H2020 – SC1-FA-DTS – 2018–2020.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 397–411, 2023.
https://doi.org/10.1007/978-3-031-34586-9_26

blow right cheek, *close eyes* (stronger than *blink*), *look left*, *look right*, *frown*, *hide lower lip*, *hide upper lip*, *kiss*, *kiss left* (moving the mouth to the left), *kiss right*, *open eyes* (more than normal), *open mouth*, *press lips*, *rise eyebrows*, *rise nose* (as if it would smell bad), *show teeth* (not smiling), *smile* (without showing teeth), *smile left* (rise left mouth corner), *smile right*, *tongue forward*, *tongue left* and *tongue right*. During an OR session, the trainee would exercise these gestures several times, starting from a neutral facial expression until the maximum gesture intensity is reached and then relaxing again.

The OR process is highly demanding in terms of expert supervision, especially when concerning elderly people [13]. Consequently, minimising the need for caregiver support during the rehabilitation process with the less possible impact on user experience is one of the main open challenges in the field to date [22]. Dialogue Systems (DS) allow human-machine natural language communication through text or speech, just as humans interact with each other. Despite their potential to automate caregiver support resembling the traditional way, their application has not been fully explored for OR. One of the main barriers to exploit DS in the rehab setting is the linguistic expertise required to model each rehabilitation process in terms of intents, entities and dialogue rules [17]. This makes it difficult for professionals to codify their knowledge in the form of dialogue. In addition, the feasibility of automated spoken interaction with mild speech impairments related to orofacial disorders has not yet been tested.

On the other hand, a smart OR system requires Facial Expression Recognition (FER) algorithms capable of efficiently spotting macro and micro expressions (i.e., gestures), together with their degree of achievement to a canonical reference in order to provide real-time feedback and evaluate the progression. This is challenging because state-of-the-art FER methods typically handle fewer facial gestures than those mentioned above. Moreover, the most accurate methods tend to have a higher complexity that might hinder their deployment in devices with limited computational resources, such as smartphones [23]. Besides, each person's neutral expression varies from person to person and, therefore, might prevent FER models from generalizing well to all facial appearances.

An additional challenge in the field of smart rehabilitation in general is the variety of smart devices such as smartphones, tablets and smart speakers that are progressively growing in use for ubiquitous rehabilitation applications [20]. In parallel, traditional client-server architectures are also transitioning towards more distributed architectures [6], demanding to minimise the transfer of sensitive data over the Internet.

In order to address the challenges described above, this work proposes a novel combination of skill-based DS and FER algorithms towards engaging AI-powered automatic OR user experience. More specifically, our contributions can be summarised as follows:

- A smart OR system architecture that allows blending the output of edge-deployed spoken interaction and computer vision modules capable of recognizing spoken keywords and orofacial expressions, with natural language interaction dialogues derived from knowledge encoded directly by rehabilitation professionals in dialogue skills.

- A spoken keyword spotting (KWS) phrase and model experimentally shown to be robust to mild speech impairment.
- A FER method to measure the degree of achievement of macro and micro facial gestures compared to a canonical reference effectively and efficiently.
- The OROFACE dataset to support further research in this field. The dataset can be downloaded from here: https://datasets.vicomtech.org/di24-oroface-dataset/oroface.zip

The remaining sections of the paper are structured as follows: Sect. 2 analyses previous work done under the scope of mixed automated rehabilitation approaches involving FER and DS; Sect. 3 introduces the proposed knowledge-driven dialogue and visual perception system architecture for smart OR; Sect. 4 evaluates this against other state-of-the-art (SOTA) prototypes and shows experimental results for the KWS and FER algorithms; finally, Sect. 5 draws upon the conclusions and potential lines for future work.

2 Related Work

2.1 Smart Rehabilitation Systems and Dialogue

Medical rehabilitation is a procedure executed by professionals including physiatrists, physiotherapists, nursing personnel, occupational therapists or diverse medics that can provide a diagnosis involving rehabilitation. Typically either patients or medical field experts need to be relocated on site for a rehabilitation session. When relocation is not an option, telehealth mechanisms flourish with the intention of performing secure virtual rehabilitation activities remotely [19]. Virtual Reality and friendly graphical user-interfaces have been added to remote rehabilitation systems to perform cognitive and physical recovery exercises supervised by a trainer [8]. Auto data gathering techniques through guided questionnaires were innitially proposed to end-users but, more recently, robotics and computer vision technologies have became more prominent to monitor and register patient performance and complete electronic health reports, which are later evaluated by competent healthcare experts [5]. However, patients still prefer to have nursing personnel next to their remote controlled rehabilitation machinery [11], highlighting the need for more natural interaction mechanisms with smart telerehabilitation systems that resemble the traditional face-to-face recovery procedure.

Dialogue Systems (DS) have been introduced in diverse knowledge fields, to provide human-like interactions with decision making capabilities independent from expert personnel involvement [14]. To prevent negative patient experience, smart rehabilitation systems shall not only collect user generated data and keep track of the exercises performed on behalf of the caregivers, but also interact as humanly as possible in order to engage individuals: informing, guiding, recommending and motivating with personalized content [4]. In this sense, speech technologies can narrow the user experience gap introducing spoken input and responses [15] as in a traditional rehabilitation process, while replacing more common but less natural visual user interfaces. A key technology in voice-based interaction is Spoken Keyword Spotting (KWS), which enables triggering attention from the system using a custom spoken word or short phrase. It also allows to initiate interactions only when users want to, enhancing user experience and acceptance [10]. Although spoken interaction is only feasible for patients with orofacial disorders not affecting speech production or leading to mild speech impairments, a comparable user experience can be achieved through chat-like text interactions in natural language. In addition, DS can fill the interactive conversational role of health-care professionals, exploiting domain knowledge in order to provide correct answers [7]. Unfortunately, the development of DS still requires considerable manual effort and expert linguistic knowledge. In order to address this problem, customisable conversation structures or dialogue skills that can adapt to each patient and conversational agent [12] have started to be exploited.

2.2 Facial Expression Recognition in Orofacial Rehabilitation

FER approaches are composed of two main phases: (1) facial image preprocessing and (2) facial expression feature classification. Image pre-processing typically includes the following stages [23]: (1) facial region detection and alignment, (2) frame normalization, and (3) motion magnification. The first stage involves extracting the facial image and landmarks from the input image, then reducing the variation in face scale and in-plane rotation. Deep Neural Networks (DNNs) currently obtain the best results for these two tasks [9]. In the second stage, meaningful frames are automatically selected from the entire gesture sequence, typically by aligning the input samples into the same number of frames through temporal interpolation [3]. The third stage usually involves manipulating the sequence transformed to the frequency domain to detect subtle motion changes of micro-expressions [16]. Finally, facial expression categories are inferred from the pre-processed facial image through a Machine Learning method (e.g., another DNN).

Orofacial rehabilitation requires recognizing more macro and micro facial gestures than those usually considered by state-of-the-art FER methods and published datasets [2]. Besides, they should be measured with respect to the specific neutral expression of each person. This means that more accurate methods and specific datasets are needed. Moreover, temporal interpolations and operations in the frequency domain might be unfeasible to get real-time feedback while performing the exercises on devices with limited computational resources. Thus, for our goal, we need to deploy lightweight DNNs, and create procedures for frame normalization and motion magnification adapted to our context.

3 Proposed Approach

3.1 Architecture Design and Workflow

The proposed Smart OR System Architecture is illustrated in Fig. 1. It entails a DS as responsible for the follow-up process, relegating the specialist to a supervis-

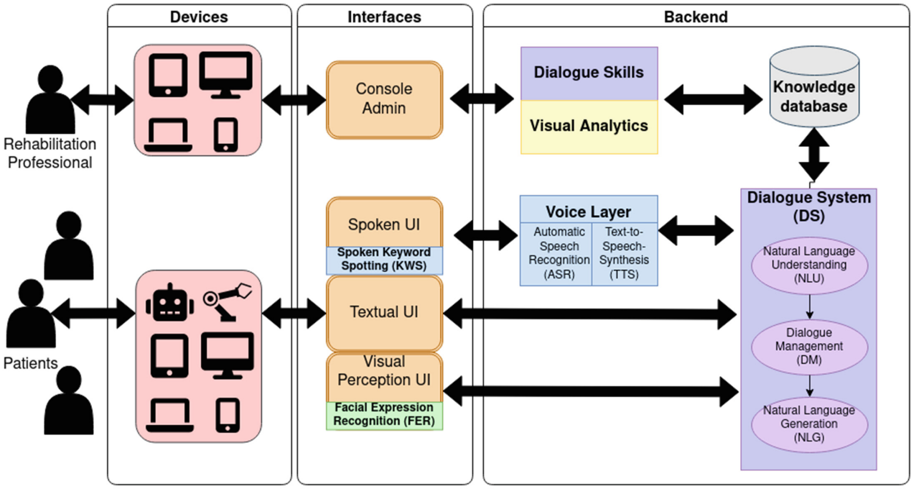

Devices
Interfaces
Backend
Rehabilitation Professional
Patients
Console Admin
Spoken UI
Spoken Keyword Spotting (KWS)
Textual UI
Visual Perception UI
Facial Expression Recognition (FER)
Dialogue Skills
Visual Analytics
Knowledge database
Voice Layer
Automatic Speech Recognition (ASR)
Text-to-Speech-Synthesis (TTS)
Dialogue System (DS)
Natural Language Understanding (NLU)
Dialogue Management (DM)
Natural Language Generation (NLG)

Fig. 1. Smart OR System Architecture

ing role consisting on introducing clinical knowledge through dialogue templates or skills to guide natural language interaction with the patient. This opposes from previous approaches that had professionals driving the whole process through screens with limited operational flexibility and full attention demand on the task. The underlying assumption is that the DS shall be capable of drawing more attention from the user than filling forms, while keeping patients active and engaged improving user experience.

The clinical knowledge encoded in the dialogue skills is stored in a knowledge database, which is then used to automatically instantiate the following DS modules:

- A Natural Language Understanding (NLU) module that performs intent classification and entity extraction on natural language input from the user.
- A Dialogue Management (DM) module that decides which is the next state in the dialogue and stores input information that may influence decision making in the dialogue memory.
- A Natural Language Generation (NLG) module that creates the appropriate natural language response to be returned by the system.

The DS supports both, text and spoken interactions through an optional voice layer that includes automatic speech recognition and speech synthesis technologies, plus a spoken keyword spotting (KWS) module. This way, OR patients without or with mild speech problems are able to interact with the system as they would do with their rehabilitation caregivers.

FER algorithms provide of real-time orofacial rehabilitation information which can be checked against medically accurate exercise templates in order to provide both trainees and instructors from useful feedback about the rehabilitation process. The output of the FER module is fed directly to the DM, which takes the visual perception information received into consideration for the decision making process in the next natural language dialogue interaction with the user.

In line with current trends, the architecture design contemplates the use of a wide variety of devices (e.g. computers, smartphones, tablets, smart speakers, robots, etc.). To ensure optimum performance, input devices shall include microphones prepared for beamforming, echo cancellation and noise reduction to deliver precise high quality audio and cameras with stabilised image and precise focusing to capture every gesture.

User information gathered from dialogue conversations and video capture peripherals is also stored in the knowledge database, which is subject to the terms of privacy and data protection for the sake of patients confidence. In order to minimise the amount of personal data to be stored in the knowledge database, the KWS and FER algorithms have been devised to be deployed on the edge. This way, only spoken interactions addressed to the system and FER performance results shall be collected. For added security and data protection, the system back-end could be deployed both on premise in a local server or in a restricted cloud network with limited access.

Additionally, the results obtained by completing the whole rehabilitation process can be checked by the expert on an intuitive multi-platform visual analytics component.

In practice, the proposed Smart OR System Architecture involves two separate workflows for rehabilitation professionals and patients:

- **Rehabilitation professionals**: use dialogue skills to define rehabilitation sessions for patients including e.g. the set and sequence of gesture exercises they should perform. Such expert knowledge is then exploited to automatically instantiate DS that guide and interact with the users throughout the rehabilitation session using natural language. Once the sessions are completed, clinicians can consult a visual analytics panel to check how the sessions went, evaluate the results achieved and plan next rehabilitation steps.
- **Patients**: open the smart OR rehabilitation application in their preferred device (i.e. computer, smartphone, tablet, robot, etc.). Calibrate the camera of the FER module to their neutral face position. The system guides them through the rehabilitation session, proposing sequences of gesture exercises adapted to their needs and providing automatic feedback on their performance in natural language. Patients without or with mild speech impairments can even interact using their voice, just as they would with their rehabilitation caregivers. After the session is completed, patients can also receive feedback from their caregivers and a new set of exercises to perform.

3.2 Dialogue Skills and KWS for OR

The most convenient method for the automatic generation of dialog rules are the so-called Dialogue Skills. To this end, each Skill is provided with the ability to manage dialogues that follow specific patterns and respect a specific domain logic. For all dialogues that partially or completely follow the default dialog structure in the Skill, the interaction rules are automatically instantiated and the Dialog Manager (DM) module gets ready to be used.

OR Dialogue Skills enable interaction capabilities such as repeating, passing, changing, completing, exiting, pausing, continuing or getting additional information about rehabilitation exercises or contacting professionals for help, following a particular conversation structure. These capabilities are linked to their respective NLU semantic tags (i.e., intents and entities) defined for the OR Skill to classify patient input and allow the DM to keep track of the dialogue state by saving and updating information related e.g. to the current exercise, the current step, the exercise number, the step number and any additional requirements (if any) as dialogue attributes. Then, the expert rules created for the Skill are able to handle dialogue state changes based on the NLU semantic tags detected. In response, users are encouraged to keep going with the OR process while receiving positive feedback or being alerted if the FER module does not detect the expected performance. Once the OR Dialogue Skill is developed, professionals only need to fill in a graphical user interface including the sequence of gesture exercises to be completed by the patients.

As mentioned before, Spoken Keyword Spotting (KWS) allows enhancing user experience and acceptance of voice-based interfaces and, thus, has been included in the proposed Smart OR System Architecture. However, a KWS phrase suitable for the OR environment precises to meet certain characteristics: it should be chosen to be as language agnostic as possible avoiding the inclusion of language-specific phonemes; and it should have a length of three to four syllables in order to avoid similarity with other words in short phrases and the complexity of long phrases. For the experimental purposes in this work, the KWS phrase “Hey Nari” has been chosen following those principles. In addition, a dataset recorded by 42 speakers of different languages has been compiled for model training and testing purposes, as described in Sect. 4. Each speaker was asked to record 12 positive audio samples containing the desired phrase and 12 negative audio samples containing diverse speech, for a total set of 1008 audio samples.

3.3 Efficient Facial Expression Recognition for OR

We need to tackle the following tasks to design an appropriate FER method for our goal:

1. Build a balanced dataset with different facial appearances in the wild, performing several trials of all the required facial gestures for orofacial rehabilitation, including neutral expressions. This dataset will allow us to build the canonical reference for the trainee’s accomplishment measurements.
2. Design an effective and efficient facial image processing strategy for frame normalization and motion magnification, especially for micro gestures.
3. Design appropriate metrics for facial gesture accomplishment, tailored to each person’s neutral expression.

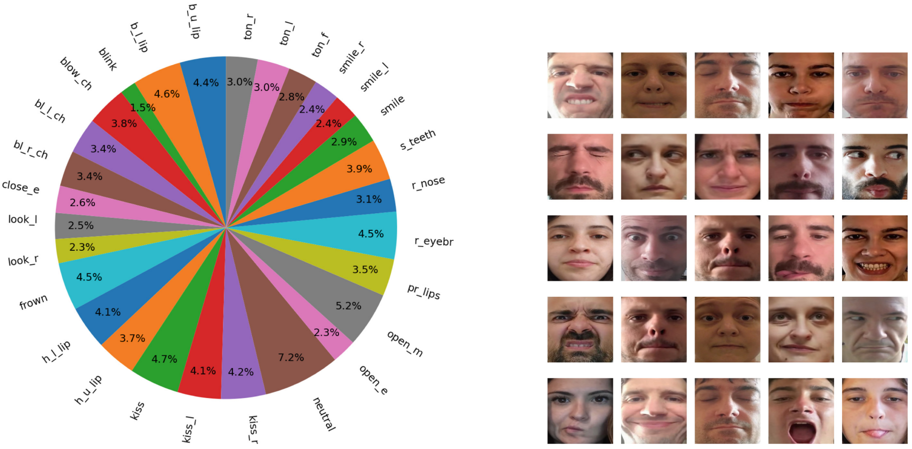


Fig. 2. Data distribution and image samples of the OROFACE dataset.

4. Analyze the visual discriminability of the required facial gestures to train an effective and efficient facial expression feature classification model.

For the first task, we first involve a professional expert in orofacial rehabilitation to define a training session. This expert is recorded from a frontal viewpoint while performing the training session, including all required trials of all the considered facial gestures. Then, similarly, other people are recorded replicating this session while watching it as guidance, as in a mirror workout session. All these people should perform all the exercises with sufficient precision to act as a further reference to others. Finally, we segment the recorded videos, extracting the frame sequences corresponding to the full gesture action, from the starting neutral expression to the ending neutral expression, but without including it. Finally, we segment separate sequences, including the neutral expression (e.g., the moment before all trials start). Following this approach, we have built the OROFACE dataset, recording 20 individuals performing the facial gestures mentioned in the introduction (28 in total) 2–3 times each, as explained above. After segmenting the videos and removing the less successful trials, the dataset contains 17,133 images, distributed as shown in Fig. 2, with the following abbreviations: b_l_lip=bite lower lip, b_u_lip=bite upper lip, blow_ch=blow cheeks, bl_l_ch=blow left cheek, bl_r_ch=blow right cheek, close_e=close eyes, look_l=look left, look_r=look right, h_l_lip=hide lower lip, h_u_lip=hide upper lip, kiss_l=kiss left, kiss_r=kiss right, open_e=open eyes, open_m=open mouth, pr_lips=press lips, r_eyeb=rise eyebrows, r_nose=rise nose, s_teeth=show teeth, smile, smile_l=smile left, smile_r=smile right, ton_f=tongue forward, ton_l=tongue left, and ton_r=tongue right.

For the second task, we propose using contrast-enhanced normalized differential images (CENDIs), computed as shown in Algorithm 1. The five input parameters are: (1) the incoming aligned facial image **I** (like the samples shown in Fig. 2), (2) an aligned facial image of the user with neutral expression of the user $\mathbf{I}_{\text{neutral}}$ obtained during an initial calibration step, (3) the grade of difference α (in the range of [0,1]), and the parameters for Contrast Limited Adaptive Histogram Equalization (CLAHE) [24] clip limit c (4) and tiles grid size g (5). CENDIs enhance the gesture's relevant areas by including the contrast between the user's actual and the neutral expression, with a small computing overhead. Figure 3 shows examples of CENDIs for different values of α.

Algorithm 1. Contrast-enhanced normalized differential image calculation.

```
procedure CalcCENDI($\mathbf{I}$, $\mathbf{I}_{\text{neutral}}$, $\alpha, c, g$)
    $\mathbf{I}_{\text{diff}} \leftarrow \mathbf{I} - \alpha \cdot \mathbf{I}_{\text{neutral}}$ (in single-precision floating-point format)
    $\text{val}_{\text{min}}, \text{val}_{\text{max}} \leftarrow \text{getMinMaxValues}(\mathbf{I}_{\text{diff}})$
    $\mathbf{I}^{\text{norm}}_{\text{diff}} \leftarrow 255 \cdot (\mathbf{I}_{\text{diff}} - \text{val}_{\text{min}})/\text{val}_{\text{max}}$
    $\mathbf{I}_{\text{HSV}} \leftarrow \text{convert2HSV}(\mathbf{I}^{\text{norm}}_{\text{diff}})$ (in 8-bit precision format)
    $\mathbf{H}, \mathbf{S}, \mathbf{V} \leftarrow \text{splitInChannels}(\mathbf{I}_{\text{HSV}})$
    $\mathbf{V}_{\text{enhanced}} \leftarrow \text{applyCLAHE}(\mathbf{V}, c, g)$ [24]
    $\text{CENDI} \leftarrow \text{convert2RGB}(\text{mergeChannels}(\mathbf{H}, \mathbf{S}, \mathbf{V}_{\text{enhanced}}))$
    return CENDI
end procedure
```

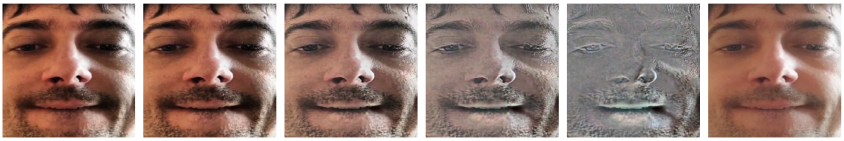


Fig. 3. Examples of CENDIs with $\alpha = 0, 0.25, 0.5, 0.75$, and 1 for the bite lower lip micro gesture, compared to the incoming aligned facial image (right).

For the third and fourth tasks, we consider using a DNN for facial expression feature classification trained with CENDIs generated from a dataset like OROFACE. Thus, the output of the DNN is a vector of scores of all the considered gestures. The closer the captured gesture's performance to the trained reference is, the higher the score for the corresponding class will be. However, even though gestures are perfectly performed, some gestures could be visually very similar (e.g., *blink* and *close eyes*), and the classifier could find difficulties in discriminating between them. Therefore, in some cases, we might require fusing some gestures, retraining, and retesting the DNN iteratively until we obtain an effective classifier, even though, in the end, we might request the user to distinguish between them for the exercises. Confusion matrices of testing data help us decide which gestures should be fused if required during this process. Nevertheless, our goal is to measure the degree of achievement to a canonical reference, and this approach is sufficient for that. In our context, we should select lightweight DNNs for this classification and facial region and landmarks detection with a good trade-off between accuracy and computational complexity for devices with limited computational resources.

4 Experiments and Evaluation

This section includes a qualitative evaluation of the proposed Smart OR System Architecture, benchmarking it against methodologies with similar characteristics found in the literature. In addition, a quantitative evaluation of the developed KWS and FER models is also presented.

Table 1 summarizes the main features of the systems analysed, which have been chosen to address smart rehabilitation or healthcare support through dialogue and/or computer vision components. The main characteristics that have been compared across systems are: whether they use dialogue to interact with the users (DS); whether dialogue skills are exploited to facilitate the development of dialogue interfaces adapted to each user (Dialogue Skills); whether users can communicate with the system using spoken interaction (Spoken Interaction); whether computer vision algorithms are exploited to automatically monitor user performance (CV); whether they consider the use of smart interconnected devices by users and propose distributed artificial intelligent deployments (IoT Edge) with several embedded machine learning (ML) algorithms.

As it can be observed, it is hard to find a solution that combines dialogue and visual perception to address the specific problem of orofacial rehabilitation. Some approaches are based on the use of smart wearable IoT devices [20] or computer vision technology alone which are not smart nor applied to facial rehabilitation [1]. Other systems exploit DS and spoken interaction [4,15], but do not target rehabilitation applications nor blend visual perception components with the dialogue. None of the published works has proposed to apply conversational skills to facilitate the inclusion of personalized expert knowledge into system-patient interactions. Regarding smart devices deployment, some approaches reflect an IoT architecture but fail to address modern limitations by embedding lightweight ML algorithms on edge devices [4].

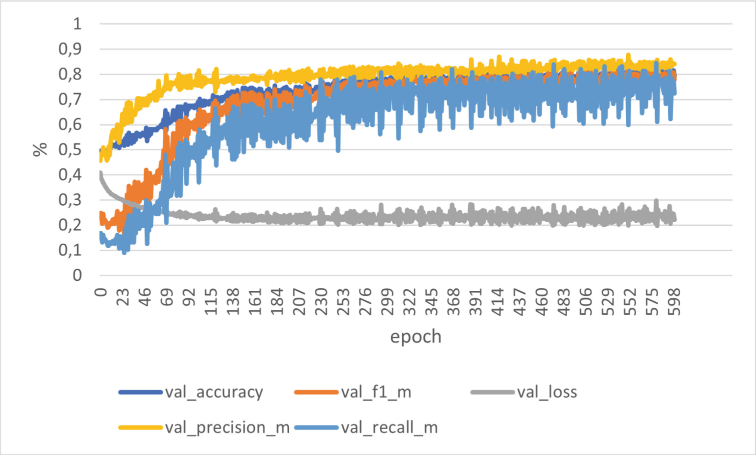
Quantitative evaluation of the KWS approach described in 3.2 has been carried out by training a model of the defined phrase on the compiled training dataset. Figure 4 shows the results achieved on the validation set for different training epochs, the best ones being those obtained with 600 epochs (final model). A small test set has been created to evaluate final model performance over different speakers and conditions as is later discussed in Table 2. As it can be seen, the maximum Accuracy obtained is 0.79, which can be considered acceptable. Overall, the Precision (0.84) of the model is higher than its Recall (0.72) leading to an F1 measure of 0.78 and 0.22 Loss.

Additionally, we have also preliminarily compared the performance of the KWS model on speech samples with and without mild speech impairment. Table 2 shows the average probabilities returned by the model for each case for a set of given test audio samples. Audio samples containing speech and noises that

Table 1. Feature Qualitative Check

Reference	**Short Description**	**Features**[a]				
		DS	***Dialogue Skills***	***Spoken Interaction***	***CV***	***IoT Edge***
[20]	Programmable device to guide rehabilitation patients	×	×	×	×	✓
[4]	Health dialog systems for patients and consumers	✓	×	✓	×	×
[15]	A dialogue monitoring scheme for a virtual doctor	✓	×	✓	×	✓
[1]	Wize Mirror - a smart, multisensory cardio-metabolic risk monitoring system	×	×	×	✓	×
Ours	Knowledge-Driven Dialogue and Visual Perception for Smart Orofacial Rehabilitation	✓	✓	✓	✓	✓

[a] Based on published research conference papers and journals.



Epoch Number	**Metrics**[a]				
	Acc	***F1***	***Loss***	***Precision***	***Recall***
600	0.79	0.78	0.22	0.84	0.72

[a] Based on micro metrics from validation phase.

Fig. 4. KWS Quantitative Analysis

differ from the KWS phrase are added to highlight the value of the results. The KWS model is inferred on overlapped fixed-size audio frames (smaller than the complete sample) trying to find a high probability value. Average probabilities reflect a clear distance between positive non-mild-speech-impairment samples (0.60) and negative samples (0.16) with positive mild-speech-impairment samples (0.45) finding their probability space somewhere in the middle of them closer to positive values. Therefore, the probability gap between those three groups appoints that this smart OR system architecture is preliminarily KWS capable on patients with mild speech impairments provided that the keyword probability detection threshold is flexibly tuned (e.g. 0.30).

Table 2. KWS Performance on Samples with and without Mild Speech Impairment

Voice conditions	Non-impaired speech	Non-impaired speech	Mild impaired speech
Wake-Up Word Audio **Sample Type**	Positive	Negative[b]	Positive
Average Probability **Achieved**[a]	0.600493349	0.157178358	0.449133078

[a] Averages are calculated based on several audio frame probabilities [0.0–1.0].

[b] Negative non-impaired speech results extrapolable to negative mild impaired speech results.

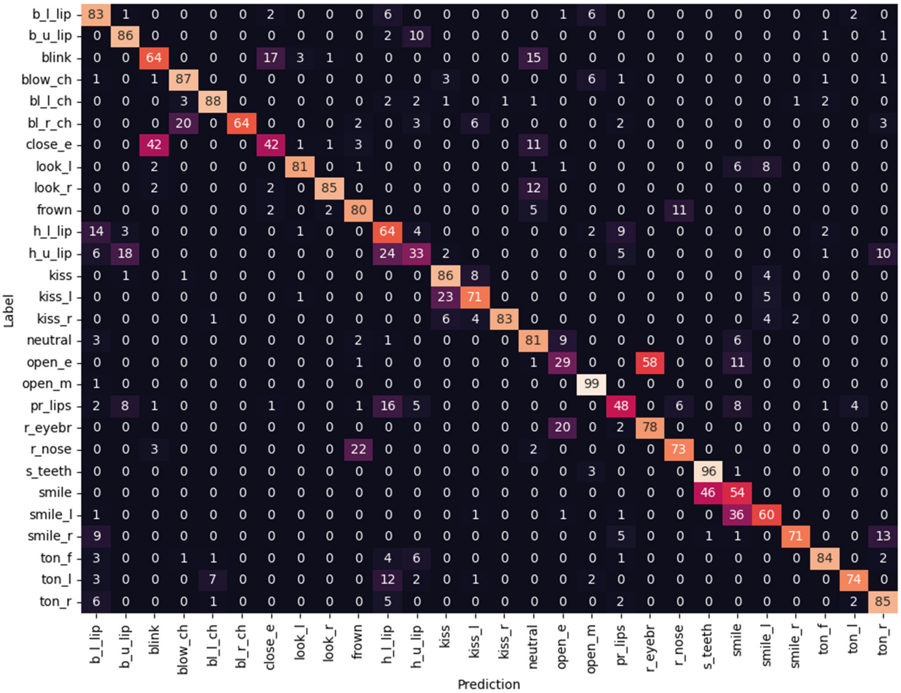
Finally, to test the convenience of CENDIs for FER in our context, we have trained ten lightweight DNNs for the first five subjects of OROFACE performing all the recorded gestures (i.e., two recognition models per subject with $\alpha = 0$ and 0.5). The data used for training each model includes the rest of the dataset's subjects, except the targeted one used for testing. We have chosen the EfficientNet-lite-0 DNN architecture [18] for these models, as it is appropriate for deploying in devices with limited computational resources.

Table 3 shows that the models trained with $\alpha = 0.5$ obtain a better recognition accuracy, as expected, because $\alpha = 0$ does not include a contrast between the user's actual and the neutral expression.

Table 3. Average gesture recognition accuracy (%) for the first five subjects of OROFACE with EfficientNet-lite-0 trained with the 28 gestures.

α **value**	**Subject 1**	**Subject 2**	**Subject 3**	**Subject 4**	**Subject 5**	**Overall**
0	75.01	66.57	68.47	63.95	57.45	**66.29**
0.5	80.83	76.53	72.58	70.78	66.78	**73.50**

In contrast, $\alpha = 0.5$ does. Figure 5 shows the normalized average confusion matrix for $\alpha = 0.5$. It reveals that in this configuration, EfficientNet-lite-0 confuses some gestures. This could be because the model is not discriminative enough, but also because in practice, some users might perform some requested gestures also moving other facial parts unconsciously (e.g., *open eyes* also rising the eyebrows, in a similar way to the *rise eyebrows* gesture).



	b_l_lip	b_u_lip	blink	blow_ch	bl_l_ch	bl_r_ch	close_e	look_l	look_r	frown	h_l_lip	h_u_lip	kiss	kiss_l	kiss_r	neutral	open_e	open_m	pr_lips	r_eyebr	r_nose	s_teeth	smile	smile_l	smile_r	ton_f	ton_l	ton_r
b_l_lip	83	1	0	0	0	0	2	0	0	0	6	0	0	0	0	0	1	6	0	0	0	0	0	0	0	0	2	0
b_u_lip	0	86	0	0	0	0	0	0	0	0	2	10	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
blink	0	0	64	0	0	0	17	3	1	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0
blow_ch	1	0	1	87	0	0	0	0	0	0	0	0	3	0	0	0	0	6	1	0	0	0	0	0	0	1	0	1
bl_l_ch	0	0	0	3	88	0	0	0	0	0	2	2	1	0	1	1	0	0	0	0	0	0	0	0	1	2	0	0
bl_r_ch	0	0	0	20	0	64	0	0	0	2	0	3	0	6	0	0	0	0	2	0	0	0	0	0	0	0	0	3
close_e	0	0	42	0	0	0	42	1	1	3	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0
look_l	0	0	2	0	0	0	0	81	0	1	0	0	0	0	0	1	1	0	0	0	0	0	6	8	0	0	0	0
look_r	0	0	2	0	0	0	2	0	85	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0
frown	0	0	0	0	0	0	2	0	2	80	0	0	0	0	0	5	0	0	0	0	11	0	0	0	0	0	0	0
h_l_lip	14	3	0	0	0	0	0	1	0	0	64	4	0	0	0	0	0	2	9	0	0	0	0	0	0	2	0	0
h_u_lip	6	18	0	0	0	0	0	0	0	0	24	33	2	0	0	0	0	0	5	0	0	0	0	0	0	1	0	10
kiss	0	1	0	1	0	0	0	0	0	0	0	0	86	8	0	0	0	0	0	0	0	0	0	4	0	0	0	0
kiss_l	0	0	0	0	0	0	0	1	0	0	0	0	23	71	0	0	0	0	0	0	0	0	0	5	0	0	0	0
kiss_r	0	0	0	0	1	0	0	0	0	0	0	0	6	4	83	0	0	0	0	0	0	0	0	4	2	0	0	0
neutral	3	0	0	0	0	0	0	0	0	2	1	0	0	0	0	81	9	0	0	0	0	0	6	0	0	0	0	0
open_e	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	29	0	0	58	0	0	11	0	0	0	0	0
open_m	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0
pr_lips	2	8	1	0	0	0	1	0	0	1	16	5	0	0	0	0	0	0	48	0	6	0	8	0	0	1	4	0
r_eyebr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	2	78	0	0	0	0	0	0	0	0
r_nose	0	0	3	0	0	0	0	0	0	22	0	0	0	0	0	2	0	0	0	0	73	0	0	0	0	0	0	0
s_teeth	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	96	1	0	0	0	0	0
smile	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	54	0	0	0	0	0
smile_l	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	36	60	0	0	0	0
smile_r	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	1	1	0	71	0	0	13
ton_f	3	0	0	1	1	0	0	0	0	0	4	6	0	0	0	0	0	0	1	0	0	0	0	0	0	84	0	2
ton_l	3	0	0	0	7	0	0	0	0	0	12	2	0	1	0	0	0	2	0	0	0	0	0	0	0	0	74	0
ton_r	6	0	0	0	1	0	0	0	0	0	5	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	85

Fig. 5. Normalized average confusion matrix for $\alpha = 0.5$, for EfficientNet-lite-0 trained with all the gestures and tested with OROFACE's first five subjects.

Moreover, these results are computed per frame in the cropped sequences, where the highest gesture intensity typically happens around the middle, and starting and ending frames might not represent the gesture properly in some cases. Thus, following the proposed approach, the problematic gestures should be fused, and then the DNN retrained and retested iteratively until sufficient accuracy is obtained.

5 Conclusions and Future Work

This paper proposes a Smart System Architecture to automate OR accurately while preserving user experience through facial expression recognition (FER) and natural language dialogue. Both, textual and spoken interactions are supported, allowing patients to communicate with the system as they would with their caregivers. In addition, dialogue skills are introduced as a mechanism to facilitate the inclusion of personalized expert professional knowledge in the system. The presented architecture also supports the use of a variety of smart devices and integrates spoken interaction and visual perception components on the edge, with the aim of minimising the transfer of sensitive data over the Internet.

The main features of the proposed Smart OR System Architecture have been benchmarked against approaches with similar characteristics found in the literature verifying that, although other published solutions use some of the same components, none of them combines conversational skills and visual perception to address the specific problem of orofacial rehabilitation. In addition, spoken keyword spotting (KWS) and FER models have also been experimentally evaluated. The developed KWS module has achieved acceptable accuracy and robustness to mild speech impairment. Regarding FER, the trained model has obtained an overall recognition accuracy of 73.50 and the OR expressions image dataset employed for experimentation is shared to support further research in the field.

Future research should further develop and confirm these initial findings by implementing a concrete use case and piloting with real users. In addition, the feasibility of using spoken interaction with a wider range of speech impairments caused by orofacial disorders should also be more thoroughly explored. The same applies to FER where further investigation should be carried out on gesture overlapping. Finally, interesting questions for future research can be derived from working towards embedding all the technological components and algorithms on smart devices on the edge.

References

1. Andreu, Y., et al.: Wize mirror - a smart, multisensory cardio-metabolic risk monitoring system. Computer Vision and Image Understanding, 148:3–22. Special issue on Assistive Computer Vision and Robotics - "Assistive Solutions for Mobility, Communication and HMI" (2016)
2. Ben, X., et al.: Video-based facial micro-expression analysis: a survey of datasets, features and algorithms. IEEE Trans. Pattern Anal. Mach. Intell. pp. 1–1 (2021)
3. Ben, X., Zhang, P., Yan, R., Yang, M., Ge, G.: Gait recognition and micro-expression recognition based on maximum margin projection with tensor representation. Neural Comput. Appl. **27**(8), 2629–2646 (2016)
4. Bickmore, T., Giorgino, T.: Methodological review: health dialog systems for patients and consumers. J. Biomed. Inform.-JBI (2021)
5. Bouteraa, Y., Abdallah, I.B., Alnowaiser, K., Ibrahim, A.: Smart solution for pain detection in remote rehabilitation. Alexandria Eng. J. **60**(4), 3485–3500 (2021)
6. Chaparro, J.D.: The shapes smart mirror approach for independent living, healthy and active ageing. Sensors, **21**(23) (2021)
7. Chen, H., Liu, X., Yin, D., Tang, J.: A survey on dialogue systems: recent advances and new frontiers. SIGKDD Explor. Newsl. **19**(2), 25–35 (2017)
8. Thumm, P.C., Giladi, N., Hausdorff, J.M., Mirelman, A.: Tele-rehabilitation with virtual reality: a case report on the simultaneous, remote training of two patients with Parkinson disease. Am. J. Phys. Med. Rehabil. **100**(5) (2021)
9. Gogic, I., Ahlberg, J., Pandzic, I.S.: Regression-based methods for face alignment: a survey. Signal Process. **178**, 107755 (2021)
10. Kepuska, V., Breitfeller, J.: Wake-up-word speech recognition application for first responder communication enhancement. In: Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense V, vol. 6201, pp, 431–438. SPIE (2006)

11. Kim, J., Lim, S., Yun, J., Kim, D.H.: Telerehabilitation needs: a bidirectional survey of health professionals and individuals with spinal cord injury in south Korea. Telemedicine Journal and e-health : the Official Journal of the American Telemedicine Association, 18(9), 713–717 (2012)
12. Liu, B., Mazumder, S.: Lifelong and continual learning dialogue systems: learning during conversation. In: Proceedings of the AAAI Conference on AI, **35**(17) (2021)
13. Maags, C.: Hybridization in china's elder care service provision. Soc. Pol. Adm. **55**(1), 113–127 (2021)
14. Major, L., Warwick, P., Rasmussen, I., Ludvigsen, S., Cook, V.: Classroom dialogue and digital technologies: a scoping review. Educ. Inf. Technol. **23**(5), 1995–2028 (2018)
15. Mallios, S., Bourbakis, N.: A dialogue monitoring scheme for a virtual doctor. In: 2015 National Aerospace and Electronics Conference (NAECON), pp. 249–253 (2015)
16. Le Ngo, A.C., Johnston, A., Phan, R.C.W., See, J.: Micro-expression motion magnification: Global lagrangian vs. local Eulerian approaches. In: 13th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 650–656. IEEE Computer Society (2018)
17. Okur, E., Sahay, S., Nachman, L.: Data augmentation with paraphrase generation and entity extraction for multimodal dialogue system (2022)
18. Tan, M., Le, Q.V.: Efficientnet: rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R., (eds.), Proceedings of the 36th International Conference on Machine Learning ICML, vol. 97 of Proceedings of Machine Learning Research, pp. 6105–6114. PMLR (2019)
19. Terrell, E.A., Bopp, A., Neville, K., Scala, D., Zebley, K.: Telerehabilitation policy report: Interprofessional policy principles and priorities. Int. J. Telerehabilitation, 13(2) (2021)
20. Tradigo, G., Vizza, P., Guzzi, P.H., Fragomeni, G., Ammendolia, A., Veltri, P.: A programmable device to guide rehabilitation patients: design, testing and data collection. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1487–1491 (2020)
21. Williams, M., Evans, P.L., Serriah, M.A.: Modern maxillofacial rehabilitation, pp. 381–420. Springer International Publishing, Cham (2022)
22. Zak, M., et al.: Frailty syndrome-fall risk and rehabilitation management aided by virtual reality (VR) technology solutions: a narrative review of the current literature. Int. J. Environ. Res. Publ. Health, **19**(5), 2985 (2022)
23. Zhou, L., Shao, X., Mao, Q.: A survey of micro-expression recognition. Image Vis. Comput. **105**, 104043 (2021)
24. Zuiderveld, K.: Contrast Limited Adaptive Histogram Equalization, pp. 474–485. Academic Press Professional Inc, USA (1994)

MM4Drone: A Multi-spectral Image and mmWave Radar Approach for Identifying Mosquito Breeding Grounds via Aerial Drones

K. T. Y. Mahima[1](✉), Malith Weerasekara[1], Kasun De Zoysa[1],
Chamath Keppitiyagama[1], Markus Flierl[2], Luca Mottola[3,4],
and Thiemo Voigt[3,4]

[1] University of Colombo, School of Computing, Colombo, Sri Lanka
{yasasm,malithk}@scorelab.org, {kasun,chamath}@ucsc.cmb.ac.lk
[2] KTH Royal Institute of Technology, Electrical Engineering and Computer Science, Stockholm, Sweden
mflierl@kth.se
[3] RISE Research Institutes of Sweden, Borås, Sweden
luca.mottola@ri.se
[4] Department of Information Technology, Uppsala University, Uppsala, Sweden
thiemo.voigt@it.uu.se

Abstract. Mosquitoes spread disases such as Dengue and Zika that affect a significant portion of the world population. One approach to hamper the spread of the disases is to identify the mosquitoes' breeding places. Recent studies use drones to detect breeding sites, due to their low cost and flexibility. In this paper, we investigate the applicability of drone-based multi-spectral imagery and mmWave radios to discover breeding habitats. Our approach is based on the detection of water bodies. We introduce our Faster R-CNN-MSWD, an extended version of the Faster R-CNN object detection network, which can be used to identify water retention areas in both urban and rural settings using multi-spectral images. We also show promising results for estimating extreme shallow water depth using drone-based multi-spectral images. Further, we present an approach to detect water with mmWave radios from drones. Finally, we emphasize the importance of fusing the data of the two sensors and outline future research directions.

Keywords: Multispectral Imagery · mmWave Radar · Aerial Drones · Object Detection

1 Introduction

Dengue and Zika are two arboviral viruses that affect a significant portion of the world population. Each year, almost 400 million dengue infections happen.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 412–426, 2023.
https://doi.org/10.1007/978-3-031-34586-9_27

Due to severe dengue fever, around half a million people each year are in need of hospitalization [39] and about 36.000 people die [23]. The number of dengue cases varies from year to year. After a reduction in many countries of the world in 2017, the numbers are increasing again [39]. In Sri Lanka alone, the number of dengue cases has been substantial in recent years with more than 150.000 cases of dengue reported in 2017 [1] (see Fig. 1). In 2017, 440 people in Sri Lanka died of dengue fever. According to government reports, the dengue patient management cost has reached 2 million USD in the year 2012 (when the number of cases was much lower than in 2017 and 2019) only for the Colombo district of Sri Lanka [24].

While there is no direct correlation between the income level of the people and the possibility of being infected by the dengue virus, the economic impact on the poor is much larger. According to Senanayake et al. [29] funds spent by households below the poverty-line for the treatment of dengue amounted to 93.7% of monthly per capita income. This is despite the fact that free health care is available in Sri Lanka.

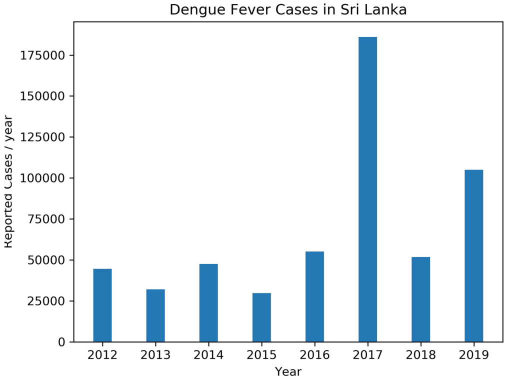


Fig. 1. Dengue Fever Cases in Sri Lanka. Some years more than 100000 cases with a strong health and economic impact, in particular on the poor part of the population.

Dengue spreads rapidly in densely populated urban areas. The principle vector species of both dengue and zika viruses are the mosquitoes Aedes aegypti and Aedes albopictus [8]. They breed in very slow-flowing or standing water pools. It is important to reduce and control such potential breeding grounds to contain the spread of these diseases. The roofs of buildings in urban environments, especially blocked gutters, provide ideal breeding grounds for Aedes. In Sri Lanka, there is a National Dengue Control Unit (see http://www.dengue.health.gov.lk/) to address this problem. Public health officials, police and military personnel visually inspect lands and buildings to locate potential mosquito breeding sites. This is difficult for the roofs of tall buildings despite that these may contain potential water collecting structures.

In this paper, we present our approach to fight dengue fever. In particular, we propose to use drones equipped with multi-spectral imagery cameras and mmWave radios to provide aerial inspection capabilities. This paper describes our system design and presents initial results in two of the projects' direction. First, we discuss how we use multi-spectral imagery to detect water from drone flights. In particular, we present our experiments to detect water retention areas from deep learning based object detection utilizing drone-based multi-spectral images. To the best of our knowledge, this is the first work that introduces a water detection method via deep-learning-based object detection and multi-spectral imagery. Moreover, estimating water depth using the bathymetric log-ratio algorithm [33] with the drone-based multi-spectral images is also not assessed yet. In summary, the main contributions of this work are as follows: (a) Demonstrate the applicability of the mmWave radios to detect water. (b) Introduce a deep-learning-based object detection network to detect water bodies via multi-spectral images. (c) Use multi-spectral images recorded by a drone to illustrate how the bathymetric log-ratio approach can be used to assess water depth.

The remainder of this paper is organized as follows: The state-of-the-art methods for drone-based detection of water using mmWave radios and multi-spectral images are outlined in Sect. 2. Section 3 discusses the system that the authors intend to develop. Section 4 presents our approaches for multi-spectral images, the related experiments and results to detect water retention areas. Section 5 discusses our drone-based water detection method using mmWave radar. Finally, Sect. 6 summarises our findings and concludes the paper.

2 Related Work

Joshi and Miller review machine learning techniques for mosquito control [17]. Like us, they focus on urban environments due to the high number of cases of mosquito-borne diseases in such areas. They highlight the challenges and progress in the area of visual detection for identifying mosquitoes. Vasconcelos et al. present an IoT-based prototype for counting mosquitoes [37]. In particular, they detect and classify mosquitoes based on the sound of their wingbeats.

Texas Instruments have presented a demonstration on applying mmWave radios to classify water and ground in a lab environment [15]. Shui et al. recently presented a system for measuring water depth using mmWave radios as we do [30]. We are aiming at going one step beyond by trying to measure water depth from drones which causes additional challenges. Other related applications of mmWave radios include the 2D rotor orbit of rotating machinery [11] as well as robust indoor mapping even in harsh environments such as in smoke-filled conditions [18].

Drones are capable of reaching locations that humans are unable to easily reach and they enable rapid observation of the ground with low operation cost. In recent research, drones are being used extensively in mosquito breeding habitat observation and other control measures such as spraying larvicides [2–5,7,9,27, 36,38]. In particular, drone-based multi-spectral imagery has also been used in several studies to determine areas that are likely to be breeding grounds [6, 22,28,32]. These studies have focused on locating relatively large water bodies in rural and peri-urban areas, such as ponds, temporary water pools and road puddles. However, we are looking for water retention areas in all urban (e.g. water retention areas on rooftops), peri-urban and rural areas.

3 System Description

This section briefly describes the system that we are implementing. Our goal is to detect the breeding places, i.e., still-standing water with mosquito larvae, in densely populated areas using drones. The detection of the breeding places happens in two steps: first drones are sent on what we call scanning flights at high altitude (around 300 m) to identify areas that need to be more closely investigated. The scanning flights will be based on digital maps that indicate potential breeding places, using open formats such as OpenStreetMap Keyhole Markup Language (KML) that facilitate the exchange of map information among involved stakeholders.

We construct the initial version of the maps with the help of public health instructors who currently do this job manually and hence have in-depth knowledge. We then automatically update the maps with data from new flights as well as weather information, for example, to include the effects of recent rainfalls that may create new potential breeding places. Based on the updated maps we construct the paths that consist of the waypoints, i.e., the potential breeding places for closer inspection flights.

In the second step, drones visit the waypoints. When arriving at a potential breeding place, the task of the drone is to detect and analyze the water area and determine whether or not it contains mosquito larvae. We investigate two approaches to solve this problem: First, we employ mmWave radios to detect water retention areas as potential mosquito habitats. Second, we use multi-spectral images to analyze the water area, measure the depth of the water and understand the larvae density. After that, we fuse the results for the final classification of the water area. Once we have detected a breeding place with mosquito larvae, the public health authorities and building owners are informed to ensure removal of the breeding place. Another option is to use spray larvicides or drop larvicide tablets into water with larvae.

4 Using Multi-spectral Imagery to Detect Mosquito Breeding Places

In this section, we discuss the usability of multi-spectral imagery to detect larval habitats. First, we discuss the drone-based multi-spectral image data that is available for processing. Then we focus on the detection of water as a potential breeding place. Here, the urban scenario is of particular interest. Finally, we emphasize the importance of water depth for larval habitats and its estimation from multi-spectral drone imagery.

4.1 Drone-Based Multi-spectral Image Data

Previous research on mosquito breeding ground detection is based on identifying near standing water bodies or water retention areas using natural colour (RGB) aerial imagery [2–4,27]. However, the information attainable from RGB images is limited when compared to that of multi-spectral imagery. For example, multi-spectral images from drones have successfully been used in combination with machine learning (ML) techniques to detect larval habitats in rural areas more accurately [6]. Our focus is on urban areas and the use of deep learning (DL) techniques to detect actual larval habitats in a challenging urban environment. Therefore, we collect our data with a MicaSense RedEdge-MX multi-spectral camera fitted onto a DJI Phantom 4 drone as shown in Fig. 2. The sensor has five spectral bands: Blue, Green, Red, Red Edge, and Near-Infrared (NIR).

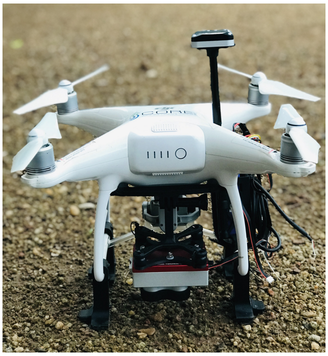


Fig. 2. MicaSense RedEdge-MX camera mounted on a DJI Phantom 4 drone.

4.2 Detecting Water Using Multi-spectral Imagery

Several studies have been done to identify water bodies using multi-spectral image datasets from satellites such as Landsat[1]. In recent years, drone-based multi-spectral images have been widely collected for different purposes. The applications range from agricultural data analysis to the detection of water areas. For this purpose, different methods have been developed.

Multi-spectral Indices

The basic methods utilize multi-spectral indices to detect water areas. We have assessed the applicability of the Normalized Difference Water Index (NDWI) [21] to classify pixels as water or non-water pixels. We have determined the

---

[1] https://landsat.gsfc.nasa.gov/data/.

index from the source images available for the spectral bands of the MicaSense RedEdge-MX sensor. The NDWI is defined as

$$\mathrm{NDWI} = \frac{\mathrm{Green} - \mathrm{NIR}}{\mathrm{Green} + \mathrm{NIR}}. \tag{1}$$

The index ranges from –1 to 1. Values above zero indicate water features. Values below or equal to zero suggest non-water features such as soil and vegetation [21].

An experiment in an environment with water on concrete ground is instructive. We learn that the NDWI is not able to properly segment the concrete area retaining water. Further, the definition of the NDWI in Eq. 1 indicates that it is highly correlated with plant water content by using NIR and Green bands [21]. Hence, using only the NDWI for identifying potential mosquito breeding places in urban environments such as water retention areas on rooftops is challenging. Therefore, more advanced methods like ML techniques have to be used for multi-spectral imagery to identify potential mosquito breeding places in urban areas.

Deep-Learning-Based Methods for Detection of Water Areas
Minakshi et al. [22] recently demonstrated the suitability of CNN-based object detection for aerial imagery by experimenting with an Inception V2 [35] network for feature extraction and a Faster Region-based CNN (Faster R-CNN) [25] with a bounding box based method to localize the areas of larval habitats. Since our primary goal is to identify potential breeding habitats in urban environments, such an object detection approach appears appealing. Here, the open question is the adequate size of the utilized bounding box. In urban environments, the diversity of water retention areas is high. Segmentation methods may become more challenging and bounding box-based detection may become more efficient and reliable.

In this study, we extend the CNN-based object detection approach proposed by Minakshi et al. [22] to process multispectral images. In order to handle 5-band multi-spectral stacked images, we modify the initial Keras Faster R-CNN network[2] as well as the pre-processing workflow. For feature extraction, we utilize either ResNet-50 [12] or VGG [31] networks. Figure 3 depicts the proposed Faster R-CNN training pipeline with the stacked multi-spectral image. We refer to it as the Faster R-CNN Multi-Spectral Water Detection (Faster R-CNN-MSWD) network.

For the experiments, we gather a multi-spectral image dataset using our camera and drone. First, we create the stacked image of 5 bands and the corresponding RGB images for all images in our dataset. Then, we use the RGB images to annotate manually the water retention regions via rectangular bounding boxes. As the size of RGB and stacked images match, we can use the bounding boxes to train the network with the stacked images. Currently, our dataset includes 112 stacked images that depict water retention areas. Finally, we use 70% of the stacked images for training and 30% for testing our Faster R-CNN-MSWD network.

---

[2] https://github.com/you359/Keras-FasterRCNN.

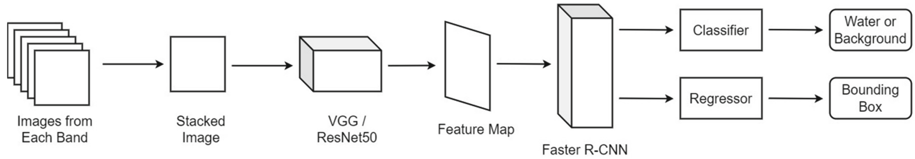


Fig. 3. Faster R-CNN model for multi-spectral images. First, the images from each band are combined into a single stacked image. A VGG or ResNet50 network is used to extract feature maps. These feature maps are then used by the Faster R-CNN to localize water areas.

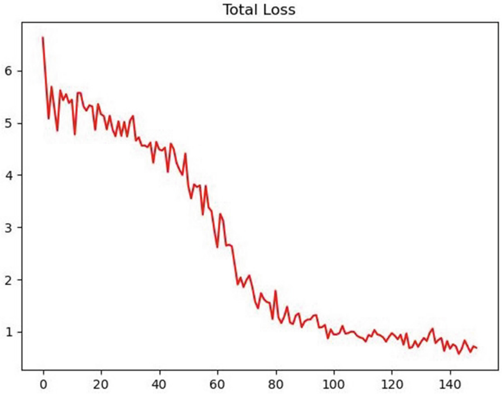


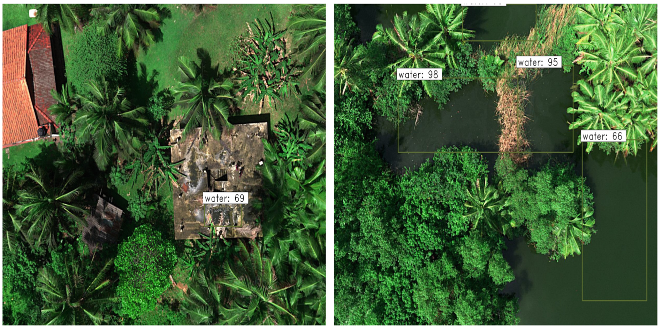
Fig. 4. Total loss of our Faster R-CNN-MSWD network with a VGG backbone. The x-axis gives the number of epochs. The y-axis shows the loss value.

We train our Faster R-CNN-MSWD with a VGG region proposal network (RPN). With this VGG backbone network, we have 136,699,171 trainable parameters. In the training phase, we reduce the total training loss to 0.575. For our test data, we achieve a mean average precision (mAP) of 0.89 at an IoU of 0.25 (Intersection over Union). However, due to the lack of training samples, we observe a relatively high number of false negatives (FN), i.e., not detected bounding boxes. The training loss curve of our Faster R-CNN-MSWD with a VGG backbone is shown in Fig. 4. Loss values for RPN and detector networks are summarized in Table 1.

Table 1. Loss values of RPN and detector networks of the Faster R-CNN-MSWD with a VGG.

RPN Classification Loss	RPN Regression Loss	Detector Classification Loss	Detector Regression Loss
0.379	0.079	0.061	0.054

Faster R-CNN-MSWD and VGG are trained using an Intel(R) Core(TM) i7-8700K CPU, an NVIDIA GTX 1080 Ti GPU and 32 GB RAM. To visualize the detection results for a multi-spectral stacked image, we add the predicted bounding boxes to the corresponding RGB image, as shown in Fig. 5.



(a) Urban Area (b) Rural Area

Fig. 5. Water retention areas as detected by the Faster R-CNN-MSWD with VGG. The bounding boxes are added to the corresponding RGB image of the given stacked multi-spectral image. Our network is able to detect water retention areas in both urban (including rooftops as shown in Fig. 5a) and rural (including large water bodies as shown in Fig. 5b) areas.

4.3 Water Depth Estimation Using Multi-spectral Imagery

The depth of water has been identified as a vital factor that influences mosquito larval development [26,32,34]. Several studies have been conducted in order to determine water depth using multi-spectral satellite imagery [10,19,33]. Recently, Sarira et al. conducted a study to determine the minimum water depth such that water areas can be accurately identified by multi-spectral images [28]. As a result, they found that there is a considerable statistical dependency between NIR reflectance and water depth. In particular, they show that it requires at least 5–10 cm depth for an accurate identification of inundated areas using NIR images.

Motivated by this, we have collected a dedicated image dataset of water buckets with varying water depths, ranging from 2–16 cm in increments of 1 cm. Our earlier paper [20] discusses our initial approach of using bathymetric models and band reflectances to estimate water depth from drone-based multi-spectral images. The log-ratio algorithm [33] in Eq. 2 has initially been introduced for satellite imagery to analyze shallow water of up to 15 m. We apply this model, determine the logarithm of the reflectance of the NIR band $R(NIR)$ and normalize it by the logarithm of the reflectance of the Blue band $R(Blue)$. The model assumes a linear relation between the depth and the log-ratio.

$$Z = m\frac{\log R(\lambda_i)}{\log R(\lambda_j)} + c \tag{2}$$

Here, Z denotes the water depth. $R(\lambda_i)$ and $R(\lambda_j)$ are the reflectance values of the NIR and Blue bands, respectively. m and c are the model parameters that can be determined by linear regression.

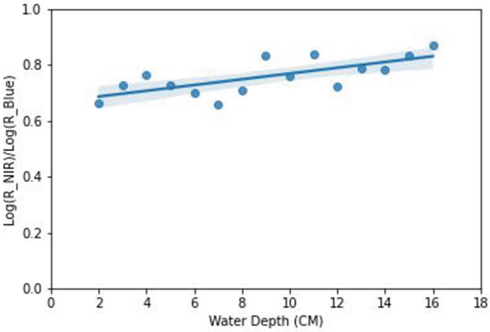


Fig. 6. Regression plot of water depth vs. log-ratio of the NIR and Blue band reflectance values. There is a linear relationship between $\log R(NIR)/\log R(Blue)$ and water depth. This indicates that we can use the bathymetric log-ratio method to estimate the water depth from drone-based multi-spectral images. (Color figure online)

Figure 6 depicts the regression plot of the initial experiment. The moderate variance of the data points around the linear regression line demonstrates the applicability of the bathymetric log-ratio algorithm to determine the depth of extremely shallow water areas using multi-spectral drone images. Note that some data points are rather noisy. In the future, we plan to improve the quality of the data in order to measure the depth of extremely shallow water areas more accurately.

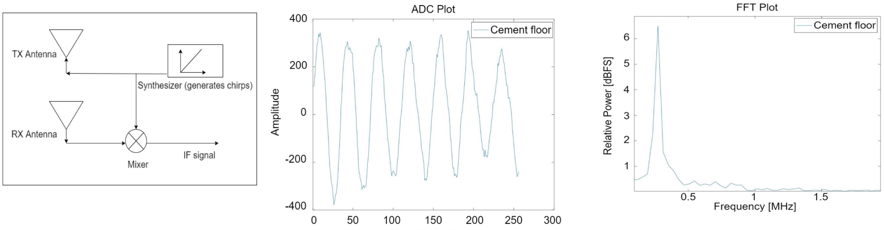
The depth of water areas is just one feature to detect larval breeding grounds more reliably. Deep-learning-based approaches are promising when identifying potential mosquito habitats in urban areas. Spectral indices and water depth will be valuable features for such learning-based methods. However, a large volume of annotated images is necessary for well-performing deep-learning networks. In the future, we will collect an annotated urban image dataset that will allow us to train a feature-based network for reliable detection of larval breeding grounds.

5 Using MmWave Radios to Detect Mosquito Breeding Places

In this section, we discuss the usability of mmWave radios to detect water areas. We start with a brief introduction of the mmWave radio technology. Then we report on our experimental setup, preliminary results gained from the experiments and identified challenges.

5.1 MmWave Radio Technology

A mmWave radio transmits an electromagnetic signal (a chirp) using its transmission (TX) antennas and captures the reflection of the chirp by its receiving (RX) antennas [16]. Then the mmWave radio passes the RX signal and TX signal to the mixer and an intermediate frequency (IF) signal is the output that contains the frequency difference between the TX and RX signals (Fig. 7a). This generated signal contains a single constant frequency and this frequency is proportional to the distance between the target from the mmWave sensor. When receiving RX signals from multiple objects with different distances, the resulting signal will be generated as a combination of multiple IF signals (Fig. 7b). By performing a Fast Fourier transform (FFT) one can compute the frequencies contained in the IF signal (Fig. 7c). The detected frequencies are then used to calculate the distance to the target and the receiving power of the signal [16].



(a) 1TX 1RX mmWave sensor block diagram

(b) IF signal (time domain)

(c) Frequencies of the IF signal (Frequency domain)

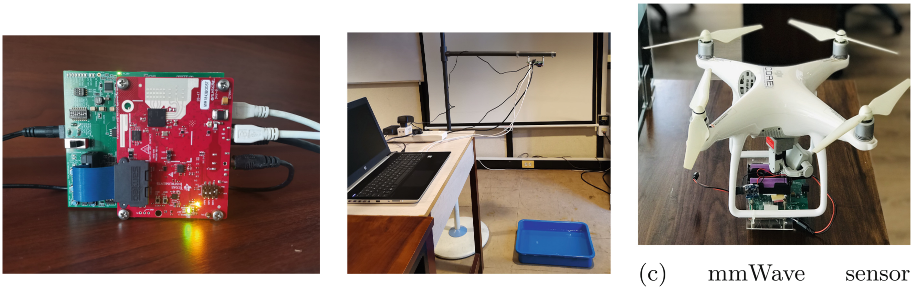
Fig. 7. mmWave Sensing

5.2 Detecting Water Using MmWave Radios

For the experiments, we use a Texas Instrument IWR1843boost mmWave sensor and Texas instrument DCA1000 evaluation module for raw data capturing (see Fig. 8a). In particular, we conduct several lab experiments to uniquely differentiate water from other target materials like soil, wood, glass pallet, copper sheet and cardboard that were placed under the sensor at a distance of 1.5 m (see Fig. 8b).

After obtaining the IF signals for each of these materials, we apply an FFT to get the corresponding frequencies and receiving power of the IF signal. Figure 9 shows that we obtain different power levels for different materials. This follows from Eq. 3 [13].

$$\text{Power Captured at RX Antenna} = \frac{P_t G_{TX} A_{RX} \sigma}{(4\pi)^3\, d^4} \tag{3}$$



(a) IWR1843boost mmWave sensor with DCA1000evm data capture card.

(b) Water detection setup in a Lab environment.

(c) mmWave sensor mounted on a DJI Phantom 4 Standard drone.

Fig. 8. Experimental setup for the mmWave sensor.

In this equation, P_t is the transmitted power, G_{TX} the TX antenna gain, A_{RX} the effective aperture area of the RX antenna, σ the radar cross-section (RCS) of the target and d is the distance.

According to Eq. 3, if we keep the target at a fixed distance of 1.5 m and the other variables are constant, the receiving power of the signal depends only on the target's RCS value. In general, the target's RCS value depends on its size, reflectivity of its surface, and its shape [14]. As a result, we receive different power levels for different materials. Hence, it is possible to only use the distance and receiving power when detecting water via mmWave radios.

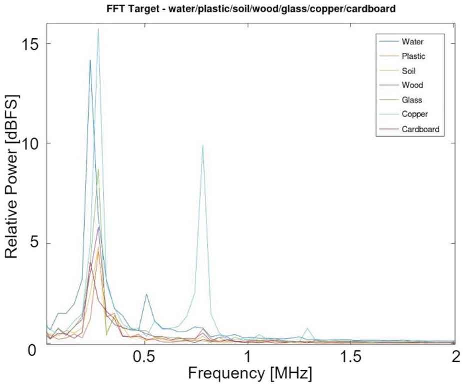


Fig. 9. Power levels of the mmWave radios for different materials. The receiving power is the highest for water but close to that of copper.

Figure 9 shows that the IF signal's receiving power for water areas is relatively high compared to other materials assessed except copper. This implies that mmWave radios are able to detect water areas. However, it also confirms that we cannot rely solely on the mmWave sensor as some materials such as copper lead to similar receiving power levels. Therefore, to identify water with very high accuracy using a second technology such as imagery is required.

5.3 Detecting Water with MmWave Radios from Drones

We integrate the IWR1843boost mmWave sensor to the DJI Phantom 4 Standard drone as depicted in Fig. 8c. Notably, the TI IWR1843boost mmWave sensor is capable of working alone without connecting to the DCA1000 evaluation module and it has its own Digital Signal Processing (DSP) chip on it. To record data from the mmWave sensor, we develop a python program[3] and run it on a Raspberry Pi Zero W module. Initially, we record two data sets targeting ground and water by hovering the drone at the same height. As depicted in the Fig. 10, the receiving power of the water areas are relatively high compared to the ground areas. This suggests that the mmWave radios can detect water from drones. The vibration of the drone generates instability in the receiving power of the signal. Hence, we use the average values for a window to stable the signal.

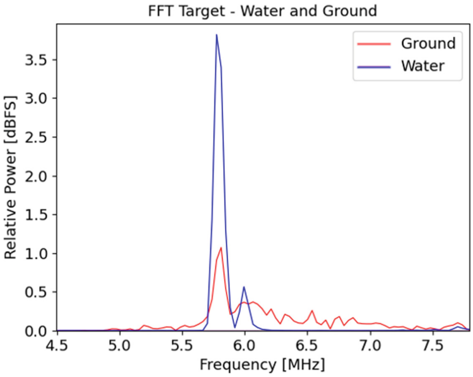


Fig. 10. Receiving Power for the ground and water

Based on the results we believe that further research on utilizing mmWave radios for detecting water using drones is essential. Moreover, we expect to verify that using mmWave radios to estimate water depth [30] is possible also from drones. Our results also indicate that fusing the results from mmWave radio and multi-spectral imagery would make the results more reliable.

6 Conclusions and Future Work

In this paper, we have evaluated multi-spectral imagery and mmWave radio waves to identify possible mosquito breeding areas by detecting water bodies.

---

[3] https://github.com/amweerasekara/mmWave-IWR1843Boost-UART-Data-Recorder.

In particular, we propose our Faster R-CNN-MSWD network that uses drone-based multi-spectral images to detect both urban and rural water retention areas. Moreover, our results show that shallow water depth can be estimated from drone-based multi-spectral images by using a bathymetric method. Our experimental results further demonstrate that drone-based mmWave radios are capable of differentiating water areas from other targeted materials. In future work, we will collect more data and improve the Faster R-CNN-MSWD network.

It is unlikely that only one method will be able to accurately identify potential mosquito breeding sites. Hence, the fusion of multi-spectral and mmWave sensor data may lead to more reliable results. A system which incorporates both approaches may be used for a future commercial drone system that is able to detect breeding sites and automatically spray larvicides or drop larvicide tablets into the detected water bodies.

Acknowledgements. This work has been partly funded by Digital Futures and the Swedish Research Council (Grant 2018-05024).

References

1. National dengue control unit, Sri Lanka. https://www.epid.gov.lk/web/index.php
2. Amarasinghe, A., Suduwella, C., Niroshan, L., Elvitigala, C., De Zoysa, K., Keppetiyagama, C.: Suppressing dengue via a drone system. In: 2017 17th International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 1–7. IEEE (2017)
3. Amarasinghe, A., Wijesuriya, V.B.: Drones vs dengue: a drone-based mosquito control system for preventing dengue. In: 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), pp. 1–6. IEEE (2020)
4. Andrade, G., et al.: Fighting back zika, chikungunya and dengue: detection of mosquito-breeding habitats using an unmanned aerial vehicle. IEEE CAS Student Design Competition 2018 (2017)
5. Bravo, D.T., et al.: Automatic detection of potential mosquito breeding sites from aerial images acquired by unmanned aerial vehicles. Comput. Environ. Urban Syst. **90**, 101692 (2021). https://doi.org/10.1016/j.compenvurbsys.2021.101692, https://www.sciencedirect.com/science/article/pii/S0198971521000995
6. Carrasco-Escobar, G., et al.: High-accuracy detection of malaria vector larval habitats using drone-based multispectral imagery. PLoS Negl. Trop. Dis. **13**(1), e0007105 (2019)
7. Case, E., Shragai, T., Harrington, L., Ren, Y., Morreale, S., Erickson, D.: Evaluation of unmanned aerial vehicles and neural networks for integrated mosquito management of Aedes albopictus (Diptera: Culicidae). J. Med. Entomol. **57**(5), 1588–1595 (2020). https://doi.org/10.1093/jme/tjaa078
8. Centers for Disease Control and Prevention: Surveillance and control of aedes aegypti and aedes albopictus in the united states. http://www.cdc.gov/chikungunya/resources/vector-control.html. Accessed 11 April 2016
9. Faraji, A., et al.: Toys or tools? utilization of unmanned aerial systems in mosquito and vector control programs. J. Econ. Entomol. **114**(5), 1896–1909 (2021). https://doi.org/10.1093/jee/toab107

10. Geyman, E.C., Maloof, A.C.: A simple method for extracting water depth from multispectral satellite imagery in regions of variable bottom type. Earth Space Sci. **6**(3), 527–537 (2019). https://doi.org/10.1029/2018EA000539
11. Guo, J., Jin, M., He, Y., Wang, W., Liu, Y.: Dancing waltz with ghosts: measuring sub-mm-level 2d rotor orbit with a single mmwave radar. In: Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021), pp. 77–92 (2021)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
13. Henriksen, S.: Radar-range equation. Proc. IEEE **63**(5), 813–814 (1975). https://doi.org/10.1109/PROC.1975.9829
14. Hess, D.W.: Introduction to RCS measurements. In: 2008 Loughborough Antennas and Propagation Conference, pp. 37–44 (2008). https://doi.org/10.1109/LAPC.2008.4516860
15. Instruments, T.: (2017). https://training.ti.com/mmwave-water-vs-ground-classification-lab
16. Iovescu, C., Rao, S.: The fundamentals of millimeter wave sensors. Texas Instruments, pp. 1–8 (2017)
17. Joshi, A., Miller, C.: Review of machine learning techniques for mosquito control in urban environments. Eco. Inform. **61**, 101241 (2021)
18. Lu, C.X., et al.: See through smoke: robust indoor mapping with low-cost mmwave radar. In: Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services, pp. 14–27 (2020)
19. Lyzenga, D.R., Malinas, N.P., Tanis, F.J.: Multispectral bathymetry using a simple physically based algorithm. IEEE Trans. Geosci. Remote Sens. **44**(8), 2251–2259 (2006). https://doi.org/10.1109/TGRS.2006.872909
20. Mahima, K.T.Y., et al.: Fighting dengue fever with aerial drones. In: International Conference on Embedded Wireless Systems and Networks (EWSN) (2022)
21. McFeeters, S.K.: The use of the normalized difference water index (NDWI) in the delineation of open water features. Int. J. Remote Sens. **17**(1996). https://doi.org/10.1080/01431169608948714
22. Minakshi, M., et al.: High-accuracy detection of malaria mosquito habitats using drone-based multispectral imagery and artificial intelligence (ai) algorithms in an agro-village peri-urban pastureland intervention site (akonyibedo) in unyama sub-county, gulu district, northern uganda. J. Public Health Epidemiol. **12**(3), 202–217 (2020). https://doi.org/10.5897/JPHE2020.1213
23. Mosquito Reviews: Statistics for mosquito-borne diseases & deaths. https://www.worldmosquitoprogram.org/en/learn/mosquito-borne-diseases/dengue (2022), https://www.worldmosquitoprogram.org. Accessed 20 May 2022
24. Thalagal, N.: Health system cost for dengue control and management in colombo district, Sri Lanka. Dengue Tool Surveilance Project, Epidemiology Unit Ministry of Health (2013)
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
26. Rohani, A., et al.: Habitat characterization and mapping of anopheles maculatus (theobald) mosquito larvae in malaria endemic areas in kuala lipis, pahang, malaysia. Southeast Asian J. Trop. Med. Public Health **41**(4), 821–30 (2010)

27. Rossi, L., Backes, A., Souza, J.: Rain gutter detection in aerial images for aedes aegypti mosquito prevention. In: Anais do XVI Workshop de Visão Computacional, pp. 1–5. SBC (2020). https://doi.org/10.5753/WVC.2020.13474
28. Sarira, T.V., Clarke, K.D., Weinstein, P., Koh, L.P., Lewis, M.M.: Rapid identification of shallow inundation for mosquito disease mitigation using drone-derived multispectral imagery. Geospat. Health **15**, 1 (2020). https://doi.org/10.4081/gh.2020.851
29. Senanayake, M., SK Jayasinghe, S., S Wijesundera, D., Manamperi, M.: Economic cost of hospitalized non-fatal paediatric dengue at the lady ridgeway hospital for children in Sri Lanka **43**(2014)
30. Shui, H., Geng, H., Li, Q., Du, L., Du, Y.: A low-power high-accuracy urban waterlogging depth sensor based on millimeter-wave FMCW radar. Sensors **22**(3), 1236 (2022)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint. arXiv:1409.1556 (2014)
32. Stanton, M.C., Kalonde, P., Zembere, K., Spaans, R.H., Jones, C.M.: The application of drones for mosquito larval habitat identification in rural environments: a practical approach for malaria control? Malar. J. **20**(1), 1–17 (2021). https://doi.org/10.1186/s12936-021-03759-2
33. Stumpf, R.P., Holderied, K., Sinclair, M.: Determination of water depth with high-resolution satellite imagery over variable bottom types. Limnology and Oceanography 48(1part2), 547–556. https://doi.org/10.4319/lo.2003.48.1_part_2.0547
34. Sutherland, A.: Mosquito management for ponds, fountains, and water gardens. UC IPM Retail Nursery & Garden Center IPM News 3 2 (2014). https://ucanr.edu/blogs/blogcore/postdetail.cfm?postnum=14396
35. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016). https://doi.org/10.1109/CVPR.2016.308
36. Valdez-Delgado, K.M., et al.: Field effectiveness of drones to identify potential aedes aegypti breeding sites in household environments from tapachula, a dengue-endemic city in southern mexico. Insects 12(8) (2021). https://doi.org/10.3390/insects12080663, https://www.mdpi.com/2075-4450/12/8/663
37. Vasconcelos, D., et al.: Counting mosquitoes in the wild: an internet of things approach. In: Proceedings of the Conference on Information Technology for Social Good, pp. 43–48 (2021)
38. Williams, G.M., Wang, Y., Suman, D.S., Unlu, I., Gaugler, R.: The development of autonomous unmanned aircraft systems for mosquito control. PLOS ONE 15(9), 1–16 (09 2020). https://doi.org/10.1371/journal.pone.0235548
39. World Mosquito Program: Statistics for mosquito-borne diseases & deaths (2022). https://mosquitoreviews.com/learn/disease-death-statistics. Accessed 20 May 2022

Machine Learning, Predictive Models and Personalised Healthcare

Spatio-Temporal Predictive Modeling for Placement of Substance Use Disorder Treatment Facilities in the Midwestern U.S

Jessica A. Pater[1](✉), Shion Guha[1,2], Rachel Pfafman[1], Connie Kerrigan[3], and Tammy Toscos[1]

[1] Parkview Research Center, Parkview Health, Fort Wayne, IN, USA
`Jessica.Pater@parkview.com`
[2] Faculty of Information and Department of Computer Science, University of Toronto, Toronto, Canada
[3] Parkview Behavioral Health, Parkview Health, Fort Wayne, IN, USA

Abstract. The inappropriate use of illegal and prescription drugs is an ongoing public health crisis across the United States and beyond. The demand for treatment services quickly outstrips the available supply, limiting access to care. Thus, a data-driven approach to assessing where new treatment facilities are to be built is an essential way to ensure new investments are strategically and optimally located. In this exploratory research, we report the findings of using 24 different public data sets to create three index variables used within a spatio-temporal modeling approach to predict what urban, suburban, and rural counties would most benefit from new substance use disorder treatments across the state of Indiana in the United States. Finally, we discuss the importance and potential limitations of taking this type of approach to develop policies that address complex societal issues.

Keywords: substance use disorder · treatment · predictive model · public policy · community health

1 Introduction

Substance use disorder (SUD) includes the continuous use of drugs and alcohol despite negative consequences. Criteria for diagnosing include the inability to stop even though you want to, neglecting responsibilities and using more of the substance than intended or using it for longer than you are meant to [11]. Approximately one in 12 adults in the U.S. has experienced a SUD in the last year, struggling with illicit drugs, alcohol, or both [12]. Indiana has the 10th worst drug problem in the nation [13] and the 14th worst overdose death rate [18]. In addition to this human tragedy, this epidemic imposes a significant burden on the healthcare system—approximately $11.3 billion annually in hospital

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 429–438, 2023.
https://doi.org/10.1007/978-3-031-34586-9_28

care for overdose clients. Healthcare systems struggle to deliver care to people needing substance use treatment with only 17% of individuals with SUD receiving any treatment in 2018 [14]. Those with SUD are at greater risk for long-term medical issues [15], and the relapse rate of 40–60% means that many cases are chronic, which positions SUD alongside other chronic illnesses [16]. Engaging clients in treatment, especially early, is critical to long-term recovery. SUDs are often multifactorial; hence, client-centered care, including shared decision-making and strong therapeutic alliance with healthcare providers, is a key strategy for effective treatment and whole-person, 360-degree care [17].

It is estimated that over 3.8 million Americans aged 12 and older receive treatment each year, representing only 8.4% of those in need [9]. Access to treatment is a major barrier to fighting this increasing epidemic. Financial/costs, stigma, geographic location, and co-occurring disorder treatment are some of the most common [10]. Geographic access is considered one of the key barriers [8]. Most treatment centers are located in urban/population settings, creating even more disparity for those in rural areas [36]. Current approaches to data-driven decisions on where to invest recovery resources use various data indicators for their assessments [37], but often lack the nuance and complexity of taking into consideration various aspects of community decay. For the purpose of this paper, we are defining community decay as a broad category of deterioration in the foundation of a specific county or region of a state. While in the strictest terms community decay refers to the break down of physical structures to a point where it is a threat to the health of a community, we look more broadly at the breakdown of social infrastructure as well.

Connected to the geographic location of treatment centers are the types of treatments offered at centers. The use of medication assisted treatment (MAT) has become the standard in long-term treatment [7] as it has been shown to reduce overall mortality and healthcare utilization [3–5]. However, it is estimated that only 41% of facilities offer one form of MAT (methadone, buprenorphine, and naltrexone) and only 3% offer all forms [2]. Geographical proximity is important as it has been found that the presence of a treatment facility is connected with decreased county-level overdose fatalities [6]. The level of care is also a further differentiation of treatment centers. Treatment facilities are broken into three levels of intensity: high, moderate, and low intensity [1].

To better understand the potential needs across different regions and counties in our state that are related to quickly evolving on-the-ground trends, we devised an exploratory analysis of various categories of social and community decay in an effort to better predict where new SUD treatment should be placed around the region to better support those in need of treatment using a spatio-temporal approach. This research makes the following contributions: connect various levels of community, social, and infrastructure decay to levels of substance use disorder in a given a specific geographic location and provide outputs of spatio-temporal predictive modeling of recommendations for a specific geographic region.

2 Related Work

There are many facets to the nature of a community in decline and/or collapse. Traditional markers include increased poverty rates [19], decreases in graduation rates [20], and increased overdose and incarceration rates [21]. Historically, various urban studies initiatives have looked at combining spatial economic and social differences to measure urban decline [22], however with the help of computational modeling approaches, these approaches can be expanded across larger regions that represent various levels of population density. Other popular theories of social decline include the reduction of in-person, social interactions. In Putnam's seminal work, he charts how declines in inter-personal community engagement maps to the rise in personal technologies as one of the factors including other aspects related to modernization [23].

There are also more unique, non-traditional ways to measure social and community decline. The Dollar Store has become a focus of interest of late related to measure of poverty and the role the store plays in trying to combat issues like food deserts [25], access to vaccines [26], and tobacco sales [24] to name a few. By creating index variables of multiple factors, it allows us to manipulate and align data, assigning weights or understanding to classes of phenomenon.

Substance use disorder has a rippling and compounding impact on individuals, families and communities [27]. There are many ways to measure the impacts of SUD on communities, including the calculations associated with loss of productivity [28], loss of life [29], and costs taken on by the healthcare systems and safety nets within the community [30,31]. Additionally, there are aspects related to impacts of substance use and abuse like erosion of trust [32], and how it relates to other social factors related to decay like decreases in high school graduation rates [33].

One way to computationally analyze the types of measures outlined above is the use of Latent class analysis (LCA). LCA is a modeling technique based on the idea that individuals can be divided into subgroups based on an unobservable construct(s) at a given point in time. Latent transition Analysis (LTA) is an extension of LCA [34]. The power of LTA is it can be used with longitudinal data, allowing to take into consideration the dynamic nature of human behavior [35]. The epidemic of SUD (including opioid use disorder) is continually evolving, thus methods like these that take nuance and complexity into consideration are critical.

3 Methods

3.1 Clinical Setting

This research was based in the state of Indiana, located in the Midwestern United States. The state is a majority rural – 70.65% of all counties are designated as rural with only 5.4% designated as urban (see Fig. 1). The total population is estimated at 6.8 million (17th most populated in the U.S.) [38]. Indiana is ranked 10th most severe state with respect to drug problems and drug related deaths, which have been consistently climbing since 2000 (see Fig. 2).

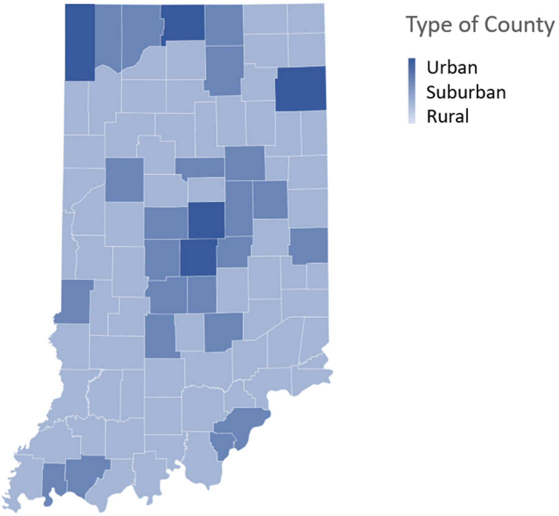


Fig. 1. Population Designation by County (Urban/Suburban/Rural) [38]

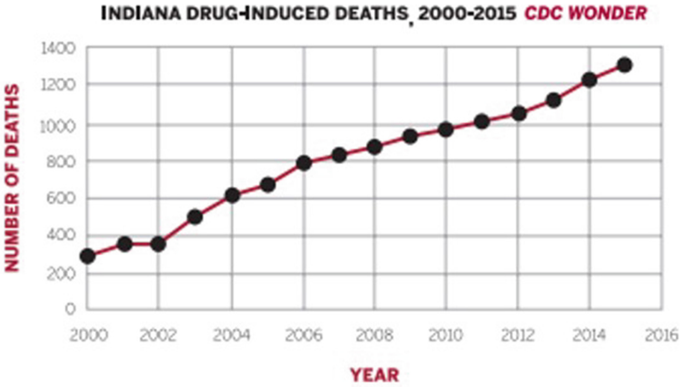


Fig. 2. Drug Induced Deaths in Indiana: 2000–2016 [18]

3.2 Data Collection

This analysis used publicly available data collected by the State and Federal governments. This is done for two key purposes - first, data collected by the government is deemed as a gold standard and is highly reliable and thus valid data for data-driven experiments. Second, publicly available data will ensure that the model can work well over a period of time and can be updated as newer iterations of collected data are updated, further reducing potential for model shift over time.

Thus, we created three index variables for this experiment: physical community decay, social community decay and addiction. The data that will comprise the initial index variables are located in the table below. Data collected for all elements were collected at the county level from 2015–2019 (See Table 1).

Table 1. Overview of data elements within each index variable and the data sources

Index Variable	Individual Data Elements	Sources
Substance Use (7)	Overdose deaths Overdose/Opioid hospitalizations Opioid prescriptions Non-fatal overdoses Fetal dependency rates Fetal death rates Proximity to current treatment	Indiana State Department of Health (ISDH) US Centers for Disease Control (CDC) US Department of Health and Human Services (DHHS)
Physical Community Decay (7)	Bankruptcy (Business) Density of low-price shopping Housing vacancy rates Job opportunities Labor turnover Housing Price Index (HPI) Foreclosure rates	US Bureau of Labor and Statistics (BLS) US Department of Justice (DOJ) US Federal Financing and Housing Agency (FHA) STATS Indiana
Social Community Decay (10)	Bankruptcy (personal) Poverty level Incarceration rates School drop out rates High school graduation rates Unemployment rates Divorce rates Child protective service placement Free/reduced lunch rates TANF (food stamps) rates	US Bureau of Labor and Statistics (BLS) US Department of Justice (DOJ) US Department of Education (DOE) Indiana Department of Child Services (IDCS) STATS Indiana ERS

3.3 Data Analysis

The analysis was conducted using a local, linear spatial regression model with multinomial outcomes for each county. Specifically, we assumed a stationary, (d+1) dimensional spatial process with a geodesic metric observed over a rectangular domain adapted from Hallin et al. 2004. We trained the resultant kernel estimator on a re-sampled training set using a standard SMOTE algorithm ($n = 1000$) using a 80-20 split. Further we used a 10-fold cross-validation to validate our results.

4 Results

Data from all 92 counties across the time period were collected and run through model. The outcomes of the model are presented here for the top three in each category of urban, rural and suburban locations. On an average, we achieved 78% accuracy with a 82% precision rate. The social and SUD related variables had more of an impact on predictions when compared to the community/physical variables (see Table 2). The demographics of the counties are the most influential as they correlate with levels of SUD recovery support in the model. Rural counties were more influenced by indicators of SUD, suburban counties were more influenced by the social decay factors and urban counties were consistently more influenced by physical decay factors.

Table 2. Outcomes of predictive model on where new SUD treatment facilities are needed based on county-level data. SD=Social Decay; PD=Physical Decay; SU=Substance Use

Type of Region	County	Influential Indicators	Model Accuracy Rate
Urban	1. Marion 2. Allen 3. Vanderburgh	- Homeownership rate (PD) - Incarceration rate (SD) - Child welfare referral rate (SD)	77%
Suburban	1. LaPorte 2. Hancock 3. Johnson	- Unemployment rate (SD) - HS graduation rate (SD) - School dropout rate (SD)	79%
Rural	1. Crawford 2. Putnam 3. Fayette	- Opioid prescription rate (SU) - Overdose deaths (SU) - Unemployment rate (SD)	73%

Additionally, two of the suburban counties are part of the greater Indianapolis (Marion County) region, the largest population center in the state. As highlighted on Fig. 3, half of the six non-urban counties have no SUD treatment facilities as of 2021, thus bolstering the outcomes of the model. Crawford and Fayette Counties are the only regions not located within a 60-minute drive of an urban population center. Interestingly, there was one urban county in the state that did not predict a high need – Lake County – which the model output showed there was a moderate need for new SUD treatment facility development. Within the *Suburban* category, Warrack was the only county that predicted having a low need for new facilities.

5 Discussion and Limitations

Complex societal issues are at the heart of many public policies, thus having a data-driven approach to address them is advantageous. The predictive model built for this analysis shows how different facets of SUD coupled with indices of decay impact are connected with population density. By understanding the impacts different indices have on need, potential biases are removed from the policy and decision making process [39]. Additionally, by taking into consideration many different types of data in the analysis, we also hedged against the

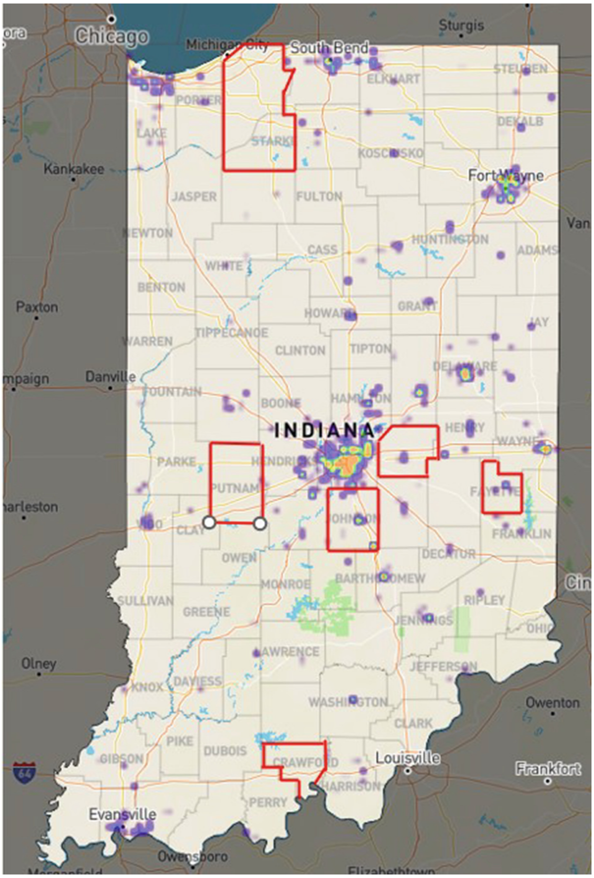


Fig. 3. Map of SUD treatment facilities in Indiana with suburban and rural outcomes of the predictive model

potential to have bias within the model outcomes because of a lack of data [40]. Making the output from the predictive model even more robust, accessible and actionable is essential. The differences in how local, state, and federal governments report various statistics is often highly variable, requiring immense levels of post-production of the data.

SUD is a rapidly evolving and ever-changing crisis. Street drugs and their complications can quickly evolve, leaving some medical complications underrecognized [41]. This coupled with the rise in prescription drug misuse [42] can leave communities scrambling as to how to get their hands around this public health issue. Utilizing computational approaches – like the one used in this analysis – on a regular, cyclical basis could provide decision makers with actionable data to inform near-term investments. Ensuring that action is taken to mitigate biases and ongoing measurements to catch potential issues as they arise are essential for the ethical integration of computing into the policy making process.

There are several limitations related to this exploratory analysis. First, computationally there are potentially biases of low populations within the model optimizations. More data and analyses would be needed to know the extent of the impact this has on the model. Additionally, we only used a 5 year time interval for this analysis. Looking at data and investments in treatment over a longer time horizon (e.g., 20 to 30 years) could provide enough window to account for other confounding impacts. While doing an analysis at a zip code level provide acceptable geographical boundaries, going even deeper to the zip code level would provide more nuance and attenuation to the results. However, due to the varying datasets used, the minimum specificity that could be obtained was to the county level. We chose to use government websites as they are perceived as gold-standard data that is more trusted and repeatedly updated than datasets curated by other special interest groups.

6 Conclusion

Data-driven decisions are critical in resource constrained environments like our state governments in the United States. Utilizing previous scholarship on the factors that go into substance use as well as social and community/physical decline of our local regions can offer unbiased and real-time information for policy makers, investors, and healthcare systems as to where investment is needed for expanded SUD treatment services. Future work on how telemedicine, mobile treatment, and other out-of-the-box treatment modalities can address the gap in access to care beyond large urban population centers and meet the last-mile healthcare needs of rural populations struggling to help individuals in their community enter into recovery.

References

1. Center for Health Policy (CHP). Treatment and Recovery for Substance Use Disorders in Indiana (2016)
2. Jones, A., Honermann, B., Sharp, A., Millett, G.: Where multiple forms of medication-assisted treatment are available. Health Affairs Blog. https://doi.org/10.1377/HBLOG20180104.835958
3. Samples, H., Williams, A.R., Crystal, S., Olfson, M.: Impact of long-term buprenorphine treatment on adverse health care outcomes in Medicaid. Health Aff. **39**, 747–755 (2020). https://doi.org/10.1377/hlthaff.2019.01085
4. Larochelle, M.R., et al.: Medication for opioid use disorder after nonfatal opioid overdose and association with mortality. Ann. Intern. Med. **169**, 137–145 (2018). https://doi.org/10.7326/M17-3107
5. Wakeman, S.E., et al.: Comparative effectiveness of different treatment pathways for opioid use disorder. JAMA Netw. Open **3**, e1920622 (2020). https://doi.org/10.1001/jamanetworkopen.2019.20622

6. Swensen, I.D.: Substance-abuse treatment and mortality. J. Public Econ. **122**, 13–30 (2015). https://doi.org/10.1016/j.jpubeco.2014.12.008
7. Volkow, N.D., Frieden, T.R., Hyde, P.S., Cha, S.S.: Medication-assisted therapies — tackling the opioid-overdose epidemic. N. Engl. J. Med. **370**, 2063–2066 (2014)
8. Substance Abuse and Mental Health Services Administration, 2020. FAQs: Provision of methadone and buprenorphine for the treatment of Opioid Use Disorder in the COVID-19 emergency. https://www.samhsa.gov/sites/default/files/faqs-for-oud-prescribing-and-dispensing.pdf
9. Center for Behavioral Health Statistics and Quality, 2017 Center for Behavioral Health Statistics and Quality 2016 National Survey on Drug Use and Health: Detailed Tables Substance Abuse and Mental Health Services Administration, Rockville, MD (2017). https://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabs-2016/NSDUH-DetTabs-2016.pdf
10. Ashford, R.D., Brown, A.M., Curtis, B.: Systemic barriers in substance use disorder treatment: a prospective qualitative study of professionals in the field. Drug Alcohol Depend. **189**, 62–69 (2018)
11. Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Association (2013). ISBN 978-0-89042-554-1. OCLC 830807378
12. Taber, M.: Programs Can Adjust to Wraparound Model. Addiction Professional (2006). https://www.hmpgloballearningnetwork.com/site/ap
13. SAMHSA. National Survey on Drug Use and Health. (2018). https://www.mobihealthnews.com/news/innovation-substance-use-disorder-treatment-5-keys-impact
14. Inc. Premier. (2019). Opioid Overdoses Costing U.S. Hospitals an Estimated $11 Billion Annually. https://premierinc.com/newsroom/press-releases/opioid-overdoses-costing-u-s-hospitals-an-estimated-11-billion-annually
15. Kiernan, J. S.: Drug Use by State: Problem Areas. WalletHub, Washington, May 2022. https://wallethub.com/edu/drug-use-by-state/35150
16. Welty, L.J., et al.: Health disparities in drug-and alcohol-use disorders: a 12-year longitudinal study of youths after detention. Am. J. Public Health **106**(5), 872–880 (2016). https://doi.org/10.2105/AJPH.2015.303032
17. National Institute on Drug Abuse. Drugs, Brains, and Behavior: The Science of Addiction: Treatment and Recovery (2018). https://nida.nih.gov/publications/drugs-brains-behavior-science-addIction/treatment-recovery
18. Center for Disease Control ADN Prevention. Drug Overdose-2020 Drug Overdose Death Rates (2020). https://www.cdc.gov/drugoverdose/deaths/2020.html
19. Albrecht, D.E., Albrecht, S.L.: Poverty in nonmetropolitan America: impacts of industrial, employment, and family structure variables. Rural. Sociol. **65**(1), 87–103 (2000)
20. Belfield, C.R., Levin, H.M.: The Return on Investment for Improving California's High School Graduation Rate. California Dropout Research Project, Santa Barbara, CA (2007)
21. Nosrati, E., Kang-Brown, J., Ash, M., McKee, M., Marmot, M., King, L.P.: Economic decline, incarceration, and mortality from drug use disorders in the USA between 1983 and 2014: an observational analysis. Lancet Public Health **4**(7), e326–e333 (2019)
22. Cheshire, P., Carbonaro, G., Hay, D.: Problems of urban decline and growth in EEC countries: or measuring degrees of Elephantness. Urban Stud. **23**(2), 131–149 (1986)

23. Putnam, R.B.: Bowling Alone: The Collapse and Revival of American Community. Simon & Schuster, New York (2000). ISBN: 978-0-7432-0301-3
24. Hall, J., Cho, H.D., Maldonado-Molina, M., George Jr, T.J., Shenkman, E.A., Salloum, R.G.: Rural-urban disparities in tobacco retail access in the southeastern United States: CVS vs. the dollar stores. Prevent. Med. Rep. **15**, 100935 (2019)
25. Chenarides, L., Cho, C., Nayga, R.M., Jr., Thomsen, M.R.: Dollar stores and food deserts. Appl. Geogr. **134**, 102497 (2021)
26. Chevalier, J.A., Schwartz, J.L., Su, Y., Williams, K.R.: EQuity Impacts Of Dollar Store Vaccine Distribution. arXiv preprint: arXiv:2104.01295 (2021)
27. Sartor, R.: The social impact of drug Aubse on community life **10**, 205–208 (1991)
28. Sorge, J.T., et al.: Estimation of the impacts of substance use on workplace productivity: a hybrid human capital and prevalence-based approach applied to Canada. Can. J. Public Health **111**(2), 202–211 (2020)
29. Naumann, R.B., et al.: Impact of a community-based naloxone distribution program on opioid overdose death rates. Drug Alcohol Depend. **204**, 107536 (2019)
30. Bahorik, A.L., Satre, D.D., Kline-Simon, A.H., Weisner, C.M., Campbell, C.I.: Alcohol, cannabis, and opioid use disorders, and disease burden in an integrated healthcare system. J. Addict. Med. **11**(1), 3 (2017)
31. Ryan, J.L., Rosa, V.R.: Healthcare cost associations of patients who use illicit drugs in Florida: a retrospective analysis. Subst. Abuse Treat. Prevent. Policy **15**(1), 1–8 (2020)
32. Hamrick, H.C., Ehlke, S.J., Davies, R.L., Higgins, J.M., Naylor, J., Kelley, M.L.: Moral injury as a mediator of the associations between sexual harassment and mental health symptoms and substance use among women veterans. J. Interpers. Viol. **37**(11–12), NP10007-NP10035 (2022)
33. Swaim, R.C., Beauvais, F., Chavez, E.L., Oetting, E.R.: The effect of school dropout rates on estimates of adolescent substance use among three racial/ethnic groups. Am. J. Public Health **87**(1), 51–55 (1997)
34. Hagenaars, J.A., McCutcheon, A.L. (eds.): Applied Latent Class Analysis. University Press, Cambridge (2002)
35. Lanza, S.T., Flaherty, B.P., Collins, L.M.: Latent Class and Latent Transition Analysis. Handbook of Psychology, pp. 663–685 (2003)
36. Pullen, E., Oser, C.: Barriers to substance abuse treatment in rural and urban communities: counselor perspectives. Subst. Use Misuse **49**(7), 891–901 (2014)
37. Congressional Research Serrvce. Location of Medication-Assisted Treatment for Opioid Addiction. In Brief. Report no. R45782., June 24 2019. https://sgp.fas.org/crs/misc/R45782.pdf
38. U.S. Census Bureau. Quick Facts: Indiana. 2020. https://www.census.gov/quickfacts/IN
39. Hutchinson, J.W., Alba, J.W., Eisenstein, E.M.: Heuristics and biases in data-based decision making: Effects of experience, training, and graphical data displays. J. Mark. Res. **47**(4), 627–642 (2010)
40. Williams, B.A., Brooks, C.F., Shmargad, Y.: How algorithms discriminate based on data they lack: challenges, solutions, and policy implications. J. Inf. Policy **8**(1), 78–115 (2018)
41. Wurcel, A.G., Merchant, E.A., Clark, R.P., Stone, D.R.: Emerging and underrecognized complications of illicit drug use. Clin. Infect. Dis. **61**(12), 1840–9 (2015). https://doi.org/10.1093/cid/civ689
42. SAMHSA. Rise in Prescription Drug Misuse and Abuse Impacting Teens, April 2022. https://www.samhsa.gov/homelessness-programs-resources/hpr-resources/rise-prescription-drug-misuse-abuse-impacting-teens

Linking Data Collected from Mobile Phones with Symptoms Level in Parkinson's Disease: Data Exploration of the mPower Study

Gent Ymeri(✉), Dario Salvi, and Carl Magnus Olsson

Internet of Things and People, Malmö University, Malmö, Sweden
gent.ymeri@mau.se

Abstract. Advancements in technology, such as smartphones and wearable devices, can be used for collecting movement data through embedded sensors. This paper focuses on linking Parkinson's Disease severity with data collected from mobile phones in the mPower study. As reference for symptoms' severity, we use the answers provided to part 2 of the standard MDS-UPDRS scale. As input variables, we use the features computed within mPower from the raw data collected during 4 phone-based activities: walking, rest, voice and finger tapping. The features are filtered in order to remove unreliable datapoints and associated to reference values. After pre-processing, 5 Machine Learning algorithms are applied for predictive analysis. We show that, notwithstanding the noise due to the data being collected in an uncontrolled manner, the regressed symptom levels are moderately to strongly correlated with the actual values (highest Pearson's correlation = 0.6). However, the high difference between the values also implies that the regressed values can not be considered as a substitute for a conventional clinical assessment (lowest mean absolute error = 5.4).

Keywords: mobile health · Parkinson's disease · mPower data

1 Introduction

Parkinson's disease (PD) is a chronic neurodegenerative disease characterised by a complex symptomatology, including impaired motor function, sleep and neuropsychiatric disorders [1]. It is the second most common neurodegenerative disease and affects more than 6 million people worldwide - a number that is expected to double in 20 years [2].

Diagnosing and assessing PD is based on clinical criteria such as the Unified Parkinson Disease Rating Scale (UPDRS). The first version of the scale was established in the 1980s, while the revised MDS-UPDRS was established in 2008 by the Movement Disorder Society (MDS) [3]. Although highly accepted

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 439–449, 2023.
https://doi.org/10.1007/978-3-031-34586-9_29

in clinical practice, these scales are used intermittently, are based on subjective criteria, and can be unreliable [4]. Subsequently, this negatively affects optimal patients' care.

Technological advancements such as smartphones and wearable devices have the potential to gather objective data for assessing disease severity on PD patients [5], thus addressing the subjectiveness of the UPDRS and MDS-UPDRS scales. To validate the usefulness of smartphones for PD treatment, the mPower observational study consists of longitudinal and frequent data collection from 14,684 individuals, both PD patients and healthy participants [6]. The study includes surveys and activities captured by smartphone sensors. Such activities include memory tests, finger tapping, vocalization test and walking test, while surveys include demographic data and other PD rating scales such as a subset of MDS-UPDRS. Previous research, cf. [7–9], has shown that these data can be used to distinguish between subjects with PD and subjects without PD, but little evidence exists that they can be associated to symptoms levels.

This paper describes an extended analysis of the mPower data set. This is done by first providing an overview to ensure that readers have a general understanding of the data set, then moves on to assessing if it is possible to predict the disease severity level based on the partial, self-reported MDS-UPDRS, together with the data collected within the smartphone-based activities.

2 Related Work

Using mobile applications to monitor health state and evaluate cognitive and motor deficits in patients with diseases that affect the central nervous system can be achieved as shown in [10]. For what regards Parkinson's Disease several studies have tried to link data from sensors and smartphones to severity levels. These include, for example, the quantification of dexterity levels of PD patients through finger tapping and spiral drawing tests using a smartphone [11]. After using machine-learning models on a set of 37 spatiotemporal features, the authors could report weak to moderate correlations between smartphone-based scores and ratings of some motor items from Part III of UPDRS. Hand tremor, another common symptom in PD, was assessed using smartphones' accelerometers in [8]. As part of that study, the authors propose an objective hand tremor severity score based on spectral power features of the acceleration signal that is shown to be significantly correlated to the self-assessed tremor score in UPDRS part II. In another study, gait analysis was conducted using an app to collect data in two 20-meter tests with PD patients walking normally and walking while performing a mental task [7]. Results from this study show how gait features such as stride time variability correlate with the UPDRS part III total score.

The mPower dataset, which we focus on in this paper, has been used in previous studies for predicting dopaminergic medication response using sensors data [9] and in the DREAM Challenge [12], where different teams competed to develop the best algorithm to e.g. differentiate between PD cases and controls. More related to the aim of our work, the mPower dataset has been also used to associate smartphone-based data with in-clinic assessments in a sub-study with 44 participants over a 6-month period [13]. The study shows how the original dataset, where volunteers were recruited online and were not followed up by any clinician, presents several biases, such as age and gender, which are not balanced when classifying PD vs non-PD. Using the controlled 44 patients cohort, authors could develop an "mPower symptom severity score" which they derived from the prediction probability of being affected by PD generated from the data of each of the activities. They showed that task performance, especially finger tapping, is predictive of self-reported PD status and correlated with in-clinic evaluation of disease severity.

While this study shows that smartphones allow remote, objective and personalized assessments of PD patients [13], the work relies on patients recruited in a controlled study. In this paper, we instead exploit the full 14-thousand uncontrolled volunteers dataset to identify links between smartphone data and symptom levels. As ground truth, we specifically use the subset of part 2 of the MDS-UPDRS self-reported questionnaire that volunteers were asked to answer.

3 Methods

3.1 The mPower Dataset

The data from mPower study was collected through Apple smartphones using ResearchKit [6]. Enrolled participants include people diagnosed with and without PD, with the latter participating as control. Patients were asked to perform 7 tasks using the smartphone: 3 surveys and 4 activity tasks. The surveys include a demographics questionnaire for PD assessment, the Parkinson Disease Questionnaire 8 (PDQ8), and a selection of items from the MDS-UPDRS, particularly questions 1.1 to 2.13, which can be self-reported and do not need clinician's observations.

The non-survey tasks consist of 4 activities: Memory activity, Tapping activity, Voice activity, and Walking activity.

1. Memory activity: used to evaluate short-term spatial memory. This is achieved by asking the participant to observe a grid of flowers that is illuminated in a sequence and to replicate the pattern in the same order by touching the flowers on the screen of the phone.
2. Tapping activity: used to measure dexterity and speed of fingers' movement. This is done by asking participants to tap on the screen of the phone with two fingers, alternatively, for 20 s.
3. Voice activity is used to measure sustained phonation by asking participants to vocalize "Aaaah" steadily for about 10 s.
4. Walking activity: used to evaluate the gait and balance of participants. They are asked to walk in a straight line for about 20 steps, turn around, stand still for 30 s and then walk again for 20 steps to get back to the same spot they started. The standing still phase also worked as a balance test.

In our study, we use the mPower data collected for the motor-related activities as input (tapping, voice, rest and walking), more concretely, the features computed and made available in [6], and the subset of part 2 of the MDS-UPDRS questionnaire, which is related to motor symptoms, as disease level reference.

Each question in the MDS-UPDRS allows one answer on a 5 levels scale, where ‘Normal’, ‘Slight’, ‘Mild’, ‘Moderate’, and ‘Severe’ are mapped to 0, 1, 2, 3, and 4 respectively. The answers thus allow us to compute a score for each part of the MDS-UPDRS and a compound one for the whole questionnaire.

In order to compare our results with existing literature such as [1], we strived to predict the full rating of part 2 of MDS-UPDRS. That part of the scale consists of 13 questions with a total score of 0 to 52. However, in mPower only 10 questions are provided (missing questions are 2.2, 2.3 and 2.11), thus reducing the maximum score to 40. As a result, we normalized the scoring by summing all the questions’ scores, dividing this score by 40 and multiplying it by 52. The formula looks as follows:

$$Normalized\_Score = \frac{\sum All\_Question\_Scores}{40} * 52 \tag{1}$$

.

3.2 Descriptive Statistics of the Data

The features computed in [6] for the 3 motor-related activities are separated into 4 subsets, because the walking activity is further split into features related to the walking phase of the activity and features related to the rest/balance phase. The number of subjects (including both PD and healthy) differs depending on the different activity data: finger tapping, walking, rest and voice (see Table 1).

Table 1. Total number of unique participants and number of tests for each activity.

	Finger tapping	Walking	Rest	Voice
Number of subjects	8003	3070	3100	5810
Number of tests	78880	34679	35407	64391

As visible in Fig. 1, the dataset is highly skewed, with few participants having performed a high number of tests and most participants having contributed with a few tests.

The number of unique participants for self-reported answers to motion-related MDS-UPDRS questions is 2024. Whereas the total number of answered questionnaires is 2305. Skewness can also be observed for these answers, with most participants (1951) having answered only once (see Fig. 3). In addition, the distribution of the score for motor symptoms computed as in Eq. 1 is also highly skewed, with most participants reporting low levels of symptoms severity (Fig. 3 and Fig. 2).

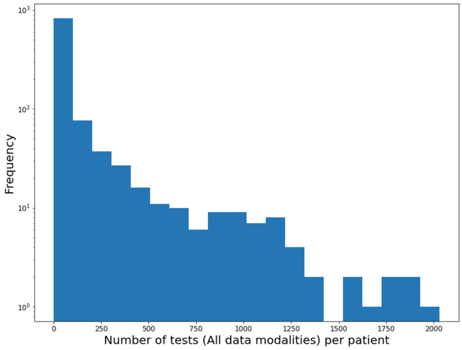


Fig. 1. The distribution of tests per patient, in logarithmic scale. Most participants contributed with a few tests.

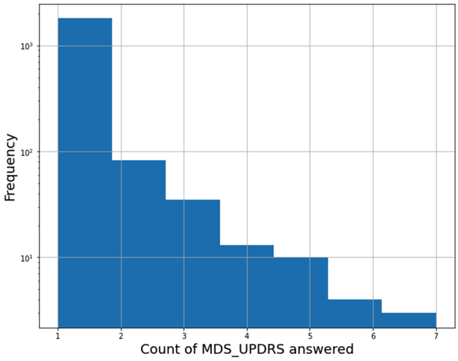


Fig. 2. Frequency of number of times the MDS_UPDRS questionnaire was answered by unique participants, in logarithmic scale. Most participants (N = 1951) answered this questionnaire only once during the study.

The skewness of these distributions is indicative of the fact that the majority of participants in the study were in relatively good health, were engaged in the study for a short time, and contributed to the study by completing a few tests and questionnaires. This was also observed in [13]. Training machine learning algorithms under these conditions is challenging and requires a well-designed pre-processing and filtering strategy.

3.3 Data Filtering and Pre-processing

As a first step, we collected the features available in the mPower dataset that corresponded to motor tasks and included additional information to link data to subjects through their 'healthCode' (unique subject identifier), 'createdOn' (timestamp of the data when it was recorded) and 'PD' (boolean variable indicating if the subject declared to have been diagnosed with PD). This yielded the following number of features for each activity type as seen in Table 2:

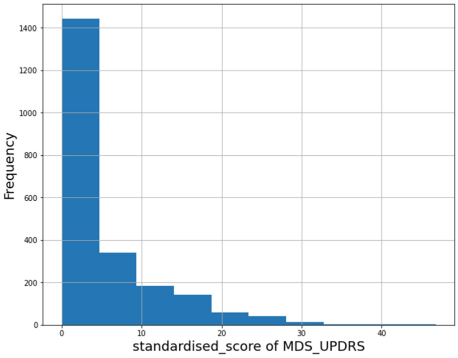


Fig. 3. Standardised score of MDS_UPDRS questions. The score is skewed towards lower values.

Table 2. The number of features for each activity type.

	Finger tapping	Walking	Rest	Voice
Nr. of features	45	116	22	16

After selecting motor features, we checked for missing values. There were missing values in some features across the different activity types, with the highest number of missing values in the ‘PD’ feature (self-declared PD diagnosis). As our analysis only focuses on actual PD patients, we decided to drop the rows where that value was missing and not try to impute them because we preferred to rely on accurate data.

In terms of motor symptoms level, we selected only those participants who answered the MDS-UPDRS questionnaire more than once. This is motivated by the risk in an uncontrolled data set like mPower - where anyone could download and use the app - that several participants wrongly declared themselves as diagnosed with PD. Without better control over the use and users, it is feasible that a number of users downloaded the app to try it, and, during that time, inserted fake data as they were testing how the app worked. Our hypothesis is, thus, that fake users would abandon the app quickly and that the data analysis would benefit from not including such users. As we observed that participants who responded the MDS-UPDRS questionnaire more than once had contributed with more tests, we used the number of answered MDS-UPDRS questionnaires as an indicator of participants’ reliability.

After selecting participants we considered reliable, we used the answers to part 2 of the questionnaire to compute the normalized score and used it as our reference symptoms level. Activity data was then associated with symptoms level, by selecting the tests within +/− 14 days from the time each MDS-UPDRS questionnaire was answered. This was based on the hypothesis that motor symptoms, while known to be fluctuating every day, would not change if averaged within a 2-week period.

In order to account for the highly skewed distribution of performed activities per patient (standard deviation of 108.89), we limited the number of activities per patient to the 50th percentile computed on the whole population (92 for Finger tapping, 97 for Walking, 80 for Rest, 85 for Voice).

All features were normalized with the PowerTransformer from the scikit-learn library. This was used to make the data more 'Gaussian-like' to minimise the skewness of the distribution of each feature [14].

Finally, the most relevant features for each activity were selected. Since we are addressing this as a regression problem, we employed a backward elimination regression technique with a linear model [15]. After feature selection, we were left with 30 features for tapping data, 60 for walking data, 11 for rest data and 12 for voice data (including meta-information such as patient ID, timestamp and reference symptoms level).

Features Collapsing Strategy. When more than one activity of the same type (finger tapping, voice, walking, rest) was found within a time window of ±14 days from the date of the reference symptoms level (derived from the answers to the MDS-UPDRS questionnaire), we computed the mean for each associated feature, grouping by participant, the symptoms level score and timestamps. This strategy has been used in previous research [13] to improve generalization and reduce the impact of identity confounding.

After averaging the features overlapping in the same time window, we merged all the features from all activities into one table together with timestamps and reference symptoms level. The resulting table contains, for each symptoms level, only one row with columns corresponding to the different features of the different activity types. Rows with missing columns, e.g. because of missing activity associated to a given symptoms level, were discarded.

Data Splitting Strategy. Similarly to [13], when splitting the data into training, test, and validation sets, we randomly shuffled the rows based on the participant identifier ('healthCode'). This was done to reduce the impact of identity confounding, which is strong in this dataset. The ratio between sets was 80/20 % between training and testing. When a validation set was required, the set was obtained from the training set by splitting the participants into 85/15 % for training and validation, respectively.

Regression Analysis. Given that our target attribute is a continuous attribute ranging between 0–52, we treated the problem of predicting the normalized symptoms level score as a regression problem. For that, we used 5 Machine Learning algorithms: Linear Regression, Lasso regression, Random Forest regressor, Support Vector Machine (SVM) and TabNet neural networks. Considering the number of data points we were left with after all the pre-processing steps, overfitting is a concern, thus, simpler models were applied such as Linear regression and Lasso regression. Furthermore, not having a huge dataset to feed, but having a high dimensional space after combining all the different data modalities, models such as SVM were supposed to perform well. TabNet was also tried as a promising, though more computationally expensive, alternative for tabular learning [16].

A Monte Carlo cross-validation was used, where we re-shuffled patients for each of the training/test and validation sets randomly 5 times [17] and then averaged the 5 results for each evaluation metric. The shuffling was done based on subject ID so that the same subject could not end in both the training and the testing set. This was done to avoid the possibility of the model to learn more about specific subjects.

Evaluation Metrics. In order to evaluate our prediction analysis of the regression models, we used the following evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Adjusted R-squared score, Pearson's correlation and Spearman's correlation.

4 Results and Discussion

The effect of pre-processing and filtering data results in the data reduction shown in Table 3. Only 293 rows were eventually used in the regression analysis. These correspond to a total of 101 participants, 80 of which were randomly chosen for the training set, and 21 for the test set. When a validation set was required, 12 patients were removed from the training set.

Table 3. Number of participants and observations per activity after each step of the pre-processing and filtering pipeline.

	Finger tapping	Walking	Rest	Voice
After dropping missing values and non-PD				
Observations	42549	23391	23998	39051
Participants	1060	640	658	968
After selecting patients with > 1 reported symptoms level and observations ±14 days apart from symptoms level				
Observations	15444	12324	12819	14756
Participants	125	102	116	124
After limiting the number of tests per patient to 50th percentile				
Observations	8305	6832	6096	7512
Participants	125	102	116	124
After averaging features				
Observations	354	295	314	352
Participants	125	102	116	124
After collapsing features				
Observations	293	293	293	293
Participants	101	101	101	101

The performances of the algorithms employed in our analysis are shown in Table 4. All the algorithms achieve similar performances, with linear regression and lasso being slightly better and showing moderate to strong correlation between predicted and regressed values (a scatter plot for the linear regression algorithm is shown in Fig. 4). The algorithm obtaining the best performance is also the simplest, linear regression, whereas Tabnet, a deep-learning algorithm suitable for datasets with a considerably higher number of rows, obtains the worst metrics.

In terms of clinical applicability, the correlation between regressed symptoms level and the reference confirms the validity of the approach (i.e. what is measured is related to motor symptoms) [18]. The metrics refer to patients who were never introduced to the algorithm before, which suggests that the algorithms generalise well. However, none of the error statistics (mean absolute error, or root mean squared error) is below the clinically significant smallest change for part 2 of UPDRS, estimated between 3.05 and 2.51 [19], which indicates that the predictions are not accurate enough.

Table 4. Evaluation metrics of the regression algorithms. The results represent the mean result of 5 random different splits.

Evaluation Metric	Linear Regression	Lasso Regression	Random Forest	SVM	TabNet
Mean Absolute Error	5.4	5.5	5.6	5.9	5.8
Root Mean Squared Error	6.6	6.7	7.0	7.1	7.5
Adjusted R-squared score	2.5	2.5	2.6	2.7	2.7
Pearson's correlation	0.6	0.5	0.3	0.4	0.2
Spearman's correlation	0.5	0.5	0.4	0.4	0.3

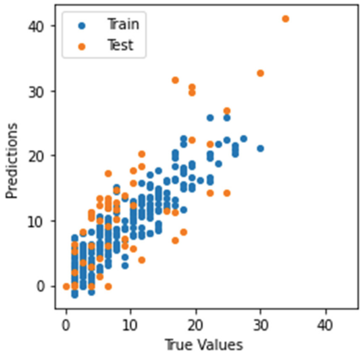


Fig. 4. Predicted symptoms level vs reference for the linear regression algorithm on the test and train sets. A positive correlation between the two quantities can be observed for the test and train set.

5 Conclusions

The mPower dataset contains an unprecedented quantity of data collected from mobile phones that can be used to detect and quantify Parkinson's disease symptoms. Given that the data was acquired in an uncontrolled manner, the dataset is skewed and likely to contain more noise. In this paper, we show how it is possible to process the dataset to focus on the parts of the dataset that is more reliable. Through such filtering, the number of participants was reduced from 1060 to 101 which we could confirm had contributed with high-quality data.

Using machine learning algorithms, we show that it is possible to correlate the data collected within the activities related to motor symptoms to the symptoms level, as measured from the answers to part 2 of the MDS-UPDRS questionnaire. The regressed level, however, still presents a high margin of error and should not be considered as a substitute for a conventional clinical assessment.

Future work could try to exploit the voluminous data available in mPower by exploring further optimization of the filtering stages with a goal to increase the number of remaining participants compared to our study and allowing more tests to be used in the training process in order to potentially improve accuracy. Additional studies could also aim at recruiting participants with a more uniform distribution across symptoms level compared with what the mPower data set currently shows, and ensuring that volunteers have been clinically diagnosed with PD.

Acknowledgment. This work was supported by the Mats Paulsson Foundation and the Internet of Things and People research center at Malmö University, funded by the Swedish Knowledge Foundation. These data were contributed by users of the Parkinson mPower mobile application as part of the mPower study developed by Sage Bionetworks and described in Synapse [doi:10.7303/syn4993293].

References

1. Sveinbjornsdottir, S.: The clinical symptoms of Parkinson's disease. J. Neurochem. **139**, 318–324 (2016)
2. Dorsey, E.R., et al.: Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016. Lancet Neurol. **17**(11), 939–953 (2018)
3. Goetz, C.G., et al.: Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov. Disord. Official J. Mov. Disord. Soc. **23**(15), 2129–2170 (2008)
4. Evers, L.J., Krijthe, J.H., Meinders, M.J., Bloem, B.R., Heskes, T.M.: Measuring Parkinson's disease over time: the real-world within-subject reliability of the MDS-UPDRS. Mov. Disord. **34**(10), 1480–1487 (2019)
5. Linares-Del Rey, M., Vela-Desojo, L., Cano-de La Cuerda, R.: Mobile phone applications in Parkinson's disease: a systematic review. Neurología (English Edition) **34**(1), 38–54 (2019)
6. Bot, B.M., et al.: The mpower study, parkinson disease mobile data collected using researchkit. Sci. Data **3**(1), 1–9 (2016)

7. Su, D., et al.: Simple smartphone-based assessment of gait characteristics in Parkinson disease: validation study. JMIR Mhealth Uhealth **9**(2), e25451 (2021)
8. Kuosmanen, E., et al.: Smartphone-based monitoring of Parkinson disease: quasi-experimental study to quantify hand tremor severity and medication effectiveness. JMIR Mhealth Uhealth **8**(11), e21543 (2020)
9. Chaibub Neto, E.L.I.A.S., et al.: Personalized hypothesis tests for detecting medication response in Parkinson disease patients using iPhone sensor data. In: Biocomputing 2016: Proceedings of the Pacific Symposium. World Scientific, 2016, pp. 273–284 (2016)
10. Lauraitis, A., Maskeliūnas, R., Damaševičius, R., Krilavičius, T.: A mobile application for smart computer-aided self-administered testing of cognition, speech, and motor impairment. Sensors **20**(11), 3236 (2020)
11. Aghanavesi, S., Nyholm, D., Senek, M., Bergquist, F., Memedi, M.: A smartphone-based system to quantify dexterity in Parkinson's disease patients. Inform. Med. Unlocked **9**, 11–17 (2017)
12. Sieberts, S.K., et al.: Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's disease digital biomarker dream challenge. NPJ Digit. Med. **4**(1), 1–12 (2021)
13. L. Omberg, E., et al.: Remote smartphone monitoring of Parkinson's disease and individual response to therapy. Nat. Biotechnol. **40**(4), 480–487 (2022)
14. Yeo, I.-K., Johnson, R.A.: A new family of power transformations to improve normality or symmetry. Biometrika **87**(4), 954–959 (2000)
15. Seabold, S., Perktold, J.: Statsmodels: econometric and statistical modeling with python. In: 9th Python in Science Conference (2010)
16. Arik, S.Ö., Pfister, T.: Tabnet: attentive interpretable tabular learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 8, pp. 6679–6687 (2021)
17. Dubitzky, W., Granzow, M., Berrar, D.P.: Fundamentals of Data Mining in Genomics and Proteomics. Springer Science & Business Media, New York (2007). https://doi.org/10.1007/978-0-387-47509-7
18. Heale, R., Twycross, A.: Validity and reliability in quantitative studies. Evid. Based Nurs. **18**(3), 66–67 (2015)
19. Horváth, K., et al.: Minimal clinically important differences for the experiences of daily living parts of movement disorder society-sponsored unified Parkinson's disease rating scale. Mov. Disord. **32**(5), 789–793 (2017)

An Exploratory Study of the Value of Vital Signs on the Short-Term Prediction of Subcutaneous Glucose Concentration in Type 1 Diabetes – The GlucoseML Study

Daphne N. Katsarou[1], Eleni I. Georga[1,2], Maria Christou[3], Stelios Tigas[3],
Costas Papaloukas[4], and Dimitrios I. Fotiadis[1,2](✉)

[1] Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, 45110 Ioannina, Greece
`fotiadis@uoi.gr`
[2] Biomedical Research Institute, FORTH, University Campus of Ioannina, 45110 Ioannina, Greece
[3] Department of Endocrinology, University Hospital of Ioannina, 45110 Ioannina, Greece
[4] Department of Biological Applications and Technology, University of Ioannina, 45100 Ioannina, Greece

Abstract. The daily self-management of type 1 diabetes (T1D) has benefitted from the advancements in real-time continuous glucose monitoring and hybrid closed-loop insulin delivery. These technologies comprise, in parallel, significant sources of data providing insight into daily glucose control and insulin treatment. The concurrent real-time continuous monitoring of vital signs, 24/7, complements the exploitable information for one individual. In this study, we investigate whether respiratory, hemodynamic, and body temperature vital signs correlate linearly with the subcutaneous glucose concentration in T1D, and improve its short-term, up to 60-min ahead, prediction as compared to a univariate model. To verify our research hypothesis, (i) we approximate the prediction of glucose concentration via a long short-term memory (LSTM) function of the recent 30-min history of glucose and those vital signs with a Pearson's $r > 0.5$, and (ii) contrast its performance with that of the univariate model. LSTM has been trained and tested individually, using a dataset with 22 T1D people monitored for 2 or 4 weeks. Our analysis identified that: (i) subcutaneous glucose concentration is linearly correlated principally with heart rate and systolic blood pressure, and (ii) the value of the vital signs lies in the improvement of the predictions in hypoglycaemia as the prediction horizon (PH) increases, where we observed a substantial reduction of erroneous predictions from 19% to 7% for a PH of 60 min.

Keywords: Type 1 Diabetes · Glucose Prediction · Machine Learning · Continuous Glucose Monitoring · Vital Signs

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 450–459, 2023.
https://doi.org/10.1007/978-3-031-34586-9_30

1 Introduction

Type 1 diabetes (T1D), as a chronic autoimmune disorder of the glucose-insulin metabolism, is characterised by an abnormal blood glucose concentration regulation leading to hyperglycaemia (defined as glucose concentration values above or equal to 180 mg/dL) [1]. The advancements in continuous glucose sensing technologies as well as insulin analogues and continuous subcutaneous insulin infusion technologies provide people living with diabetes (PLWD) and healthcare professionals with a continuous healthcare model where insulin treatment and other modifiable factors affecting diabetes management (e.g., diet, physical activity) are adjusted efficiently according to objective measures of the glucose control [2, 3]. Hyperglycaemia is an independent risk factor of long-term micro- and macro-vascular complications, which, in turn, account for increased morbidity and mortality rates in diabetes. Nonetheless, hypoglycaemia (defined as glucose concentration values below or equal to 70 mg/dL) remains a most significant barrier for PLWD since its acute life-threatening symptoms have a direct effect on the quality of their life (QoL) [4, 5].

Whilst the problem of short-term prediction of subcutaneous glucose concentration in T1D has been extensively investigated and modelled via linear or non-linear, machine-learning (ML)-based functions, provided the wealth of data amassed by the continuous glucose monitoring (CGM) systems, the meta-analysis of a number of recent systematic reviews points out that it remains still a challenging problem which calls for a more comprehensive representation of the biology, behaviour and context of PLWD within the prediction function [6, 7]. The addition to the model's input information on the insulin therapy, carbohydrates consumption, physical activity or physiological characteristics (e.g., heart rate, galvanic skin response, skin temperature) has been shown to improve the error of predictions as compared to the univariate prediction models [8, 9]. In addition, the combination of mechanistic models of the processes of intestinal absorption of carbohydrates and the absorption of subcutaneously injected insulin increases the granularity of the information fed into the prediction function [10, 11]. An interesting finding of head-to-head comparison studies is that non-linear univariate models compare with linear ones with respect to prediction horizons up to 60 min, which might be reflective of the fact that short-term glucose regulation features linear dynamics or can be attributed to the existence of unmodelled inputs [12–14]. On top of these, the clinical impact of prediction errors has been also considered as not only a performance metric but also an optimisation metric embedded into the training of the ML model [15].

In this study, we examine for the first time in the literature whether daily haemodynamic changes, as they are captured by: heart rate, heart rate variability, systolic blood pressure, diastolic blood pressure, stroke volume, systemic vascular resistance, cardiac output, cardiac index, pulse pressure, mean arterial pressure, may improve the short-term prediction of subcutaneous glucose concentration in T1D. The transient cardiac stress associated with hypoglycaemic events is well documented in the literature, while its predictive capacity has not been studied yet. To this end, we contrast the predictive capacity of a univariate long short-term memory (LSTM) neural network with that of a multivariate LSTM fed with the cardiac indices alongside CGM values. A dataset of 29 T1D patients is leveraged for this purpose generated by the GlucoseML-Phase I observational prospective study, which enables an unbiased evaluation of the prediction models.

2 Materials and Methods

2.1 Materials

The GlucoseML-Phase I prospective study has been designed as an observational study aiming at the collection of real-world data from T1D patients following an intensive insulin therapy scheme (i.e., multiple daily injections (MDI) of insulin or continuous subcutaneous insulin infusion (CSII)) (as it is shown in Table 1), which data will comprise the training/validation/test sets upon which short-term prediction models of subcutaneous glucose concentration time series will be development and internally evaluated. Participants use the GlucoMen Day CGM Menarini®[1] system and the Biobeat® wrist monitor[2], and, in parallel, they manually record the carbohydrate content of daily meals and administered insulin therapy using specially designed logs; the GlucoseML-Phase I study encompasses special training sessions on carbohydrates counting to alleviate the errors introduced in the collection of data. The target number of patients to be recruited and the target duration of the study has been set to 30 patients and 4 weeks, respectively. In total, 32 patients were recruited among whom 26 patients (82%) completed the 4-weeks monitoring period, 3 patients (9%) completed the 2-weeks monitoring period, while 3 patients (9%) drop out. Table 1 describes the overall characteristics of the GlucoseML-Phase I study cohort, while Table 2 presents the average glucose statistics observed at the end of the study according to the Ambulatory Glucose Profile report.

2.2 Methods

The prediction function of the subcutaneous glucose concentration time series has been approximated by an LSTM network, defined, fine-tuned and trained using the GlucoseML-Phase I dataset. The LSTM function maps the feature vector $x(t)$:

$$\mathrm{x(t)} = \left[v_1\left(t - h_{v_1}\right), v_1\left(t - h_{v_1} + \Delta t_{\mathrm{v}_1}\right), \ldots, v_1(t), \ldots, v_{14}\left(t - h_{v_{14}}\right), v_1\left(t - h_{v_{14}} + \Delta t_{\mathrm{v}_{14}}\right), \ldots, v_{14}(t)\right] \quad (1)$$

to the observed subcutaneous glucose concentration value over the next $\mathrm{t} + \mathrm{PH}$ minutes, $y(t + PH)$, where $\{v_i\}_{i=1}^{14}$ denotes the set of input variables (Table 3), h_{v_i} is the history window (expressed in min) specified for v_i, $\Delta t_{\mathrm{v_i}}$ is the sampling interval of v_i, and PH is the prediction horizon (expressed in min) [16]. In this study, we have considered two input cases: (i) Case 1 constitutes a univariate prediction problem relying on the assumption that the recent CGM profile suffices to predict its future short-term course, and (ii) Case 2 forms a multivariate prediction problem accounting additionally for the relationship between glycaemic excursions and changes in vital signs expressing the cardiovascular autonomic nervous system function. We have considered an equal history window $h_{v_i} = 30$ min for all input variables v_i, while the sampling interval of the

---

[1] https://glucomenday.com.

[2] https://www.bio-beat.com.

Table 1. Descriptive characteristics of the GlucoseML-Phase I study cohort

	Feature	Distribution of values
Demographics	*Gender*	Male: 62% (18) Female: 38% (11)
	Age	38 ± 12 years
Anthropometrics	*BMI*	Normal weight: 24% (7) Overweight: 41% (12) Obese: 34% (10)
	Waist circumference	Female: 86 (71–124 cm) Male: 96 ± 15 cm
T1D management	*Years since diagnosis*	0–12 years: 38% (11) 12–24 years: 35% (10) 24–36 years: 17% (5) 36–48 years: 10% (3)
	Type of insulin treatment	MDI: 66% (19) CSII: 34% (10)
	History of severe hypoglycaemia	59% (17)
	Baseline HbA1c	7.5 ± 1.0%
	Complications of T1D	Albuminuria: 10% (3) Retinopathy: 10% (3) Neuropathy: 3% (1)
Other metabolic comorbidities	*Dyslipidaemia*	52% (15)
	Central obesity	31% (9)
	Thyroid disease	24% (7)
	Hypertension	10% (3)
Lifestyle	*Smoking*	52% (15)
	Alcohol	7% (2)

Data are presented in the forms: percentage of patients% (number of patients), or mean ± standard deviation, or median (min-max).

subcutaneous glucose concentration time series is reduced to the one of physiological vital signs, i.e., 5 min, by applying the Dynamic Time Warping method ($\Delta t_{\mathrm{v_i}} = 5\,\mathrm{min}$ for all input variables v_i) [17].

The LSTM network, comprising two LSTM layers of 4 units and one dense output layer of one unit, has been trained and tested individually for each patient using the time series data segments defined by the first 70% of time points and the remaining 30% of the entire time series, respectively. All input variables have been scaled to [0, 1] over the training set. In Case 2, the linear relationship between the subcutaneous glucose concentration and each of the vital signs time series is examined using the Pearson's correlation coefficient, with only those vital signs featuring a > 0.5 correlation with the glucose concentration feeding the LSTM model. The hyperparameters relating to batch

Table 2. The Ambulatory Glucose Profile (AGP) report across all patients using CGM data over the entire period of the study.

AGP Report Variables	Mean ± Standard Deviation
Time Below Range – Very low (<54 mg/dL)	2.3 ± 2.6%
Time Below Range – Low (<70 mg/dL)	3.6 ± 2.3%
Time in Range (70–180 mg/dL)	61.4 ± 14.0%
Time Above Range – High (>180 mg/dL)	22.4 ± 6.4%
Time Above Range – Very high (>250 mg/dL)	10.3 ± 10.2%
Average Glucose (mg/dL)	159.9 ± 26.2%
Glucose Management Indicator (%)	7.1 ± 0.6%
Glucose Variability (%)	39.5 ± 6.3%

size, number of epochs, and the optimization algorithm itself are finetuned to minimise the 2-fold cross-validated Root Mean Squared Error (RMSE) computed over the training set. The grid of hyper-parameters search space is presented Table 4. The selected model was trained using Adam optimiser with a batch size equal to 16 and iterated over 100 epochs.

Table 3. The input variables space of the GlucoseML prediction function.

Variable	Measurement method	Sampling frequency (min)
Subcutaneous glucose concentration	GlucoMen Day CGM Menarini®	1 min
Respiratory rate, oxygen saturation, heart rate, heart rate variability, systolic blood pressure, diastolic blood pressure, stroke volume, systemic vascular resistance, cardiac output, cardiac index, skin temperature, pulse pressure, mean arterial pressure	Biobeat® wrist monitor	5 min

3 Results

The performance of the prediction models was assessed using: (i) two pure error metrics, i.e., the RMSE and the Mean Absolute Percentage Error (MAPE), and (ii) the Continuous Glucose Error Grid Analysis (CG-EGA) which evaluates, in parallel, the clinical impact of the errors depending on the glycaemic range the actual glucose concentration value lies in [18]. It should be noted that the results reported herein concern 22 out of 29

Table 4. The hyper-parameters search space.

Hyper-parameters	Range
Batch size	[16, 24, 32]
Number of epochs	[50, 100]
Optimiser	[‘*adam*’, ‘*Adadelta*’]

patients of the GlucoseML study; the vital signals of 7 patients were excluded due to a considerable percentage (>50%) of missing values.

Via the Pearson’s correlation analysis, heart rate and blood pressure were selected for most patients, mean arterial pressure was retained as a feature for 2 patients, while stroke volume and skin temperature were retained as features only for one patient. Table 5 and Fig. 1 describe the distribution of test RMSEs and the test MAPEs associated with 15-min, 30-min and 60-min ahead predictions of subcutaneous glucose concentration when either Case 1 or Case 2 is applied. First, we observe that both input cases yield low errors in the case of a 15-min PH, retaining the average MAPE of 30-min and 60-min PHs below 5% and 8%, respectively. Second, we observe a modest but systematic improvement of the average prediction errors in Case 2, with their interquartile range being located lower for all prediction horizons except for the MAPE for a 60-min PH, which is also captured by the narrower standard deviation achieved in Case 2. The CG-EGA (Table 6, Fig. 2) confirms that the predictive capacity of the additional vital signs becomes evident in the hypoglycaemic range for all PHs, where Case 2 reduces apparently the Erroneous Predictions in this critical range. The contribution of Case 2 in the reduction of erroneous prediction in the hyperglycaemic or euglycemic ranges become more evident for 30-min and 60-min PHs.

Table 5. Prediction errors over the test set achieved by the LSTM models.

	RMSE (mg/dL)			MAPE (%)		
	15 min	30 min	60 min	15 min	30 min	60 min
Case 1	5.7 ± 3.0	9.8 ± 5.8	15.6 ± 7.4	2.6 ± 1.4	4.7 ± 3.3	7.7 ± 4.2
	5.1 (4.7, 5.6)	7.0 (8.3, 11.6)	13.3 (11.4, 18.2)	2.2 (2.0, 2.5)	4.2 (3.1, 5.0)	6.8 (5.7, 7.5)
Case 2	5.3 ± 2.5	8.5 ± 4.3	14.8 ± 6.3	2.5 ± 1.3	4.0 ± 2.7	7.4 ± 3.2
	4.7 (4.1, 5.6)	7.7 (5.7, 9.3)	13.7 (10.7, 17.4)	2.1 (1.7, 2.5)	3.4 (2.8, 4.5)	6.8 (5.3, 8.6)

Data are presented in the form mean ± standard deviation or median (25th, 75th quartiles).

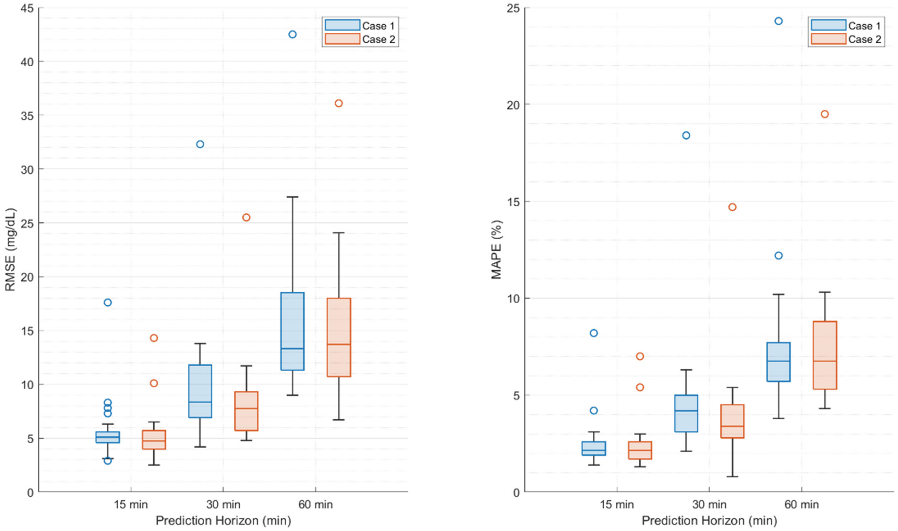


Fig. 1. Box plot of the RMSE and MAPE associated with the LSTM predictions.

Table 6. The classification of the prediction errors according to the CG-EGA.

		15 min		30 min		60 min	
		Hypo	*Hyper*	*Hypo*	*Hyper*	*Hypo*	*Hyper*
Case 1	AP	0.90 ± 0.22	0.94 ± 0.06	0.88 ± 0.14	0.91 ± 0.08	0.66 ± 0.34	0.86 ± 0.07
	BE	0.02 ± 0.03	0.04 ± 0.04	0.05 ± 0.08	0.07 ± 0.04	0.04 ± 0.08	0.07 ± 0.05
	EP	0.08 ± 0.23	0.02 ± 0.02	0.07 ± 0.09	0.03 ± 0.04	0.30 ± 0.36	0.07 ± 0.06
Case 2	AP	0.96 ± 0.05	0.94 ± 0.06	0.87 ± 0.13	0.92 ± 0.06	0.77 ± 0.19	0.86 ± 0.08
	BE	0.02 ± 0.04	0.04 ± 0.04	0.05 ± 0.07	0.06 ± 0.05	0.08 ± 0.15	0.09 ± 0.05
	EP	0.03 ± 0.03	0.01 ± 0.02	0.08 ± 0.09	0.02 ± 0.03	0.15 ± 0.17	0.05 ± 0.03

Data are presented in the form mean ± standard deviation. AP: Accurate Predictions, BE: Benign Errors, EP: Erroneous Predictions.

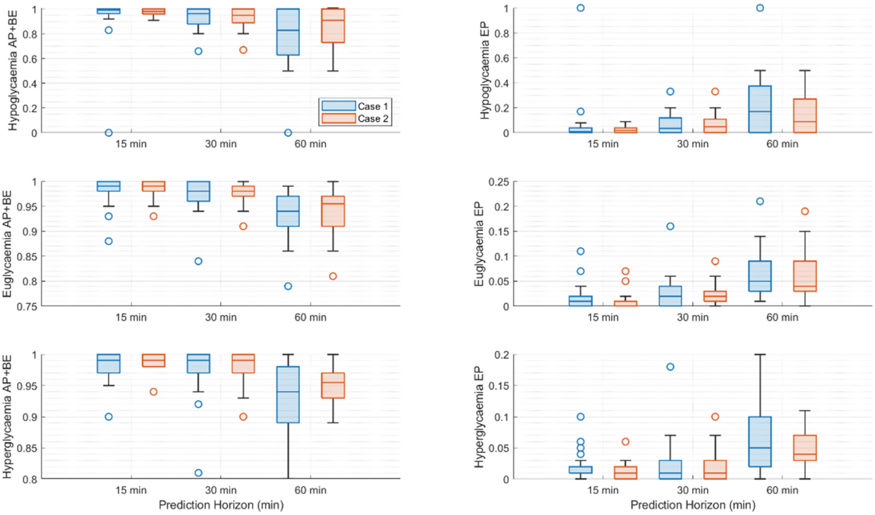


Fig. 2. Box plots of the classification of errors according to the CG-EGA. AP: Accurate Predictions, BE: Benign Errors, EP: Erroneous Predictions.

4 Discussion and Conclusions

This study explores the predictive capacity of a set of cardiac indices monitored continuously 24/7 in the context of the GlucoseML study, with respect to the short-term prediction (setting the maximum prediction horizon to 60 min) of subcutaneous glucose concentration in T1D. The initial results achieved by an LSTM network, trained and tested over a dataset of 22 people living with T1D who are monitored for a period of 2 or 4 weeks, indicate that these variables can improve the clinical accuracy of predictions in the hypoglycaemic range in spite of the modest improvements in the overall RMSEs and MAPEs.

Clinical evidence supports that hypoglycaemia brings about temporal haemodynamic changes, stimulated by the autonomic nervous system, including: *“an increase in heart rate and peripheral systolic blood pressure, a fall in central blood pressure, reduced peripheral arterial resistance (causing a widening of pulse pressure), and increased myocardial contractility, stroke volume, and cardiac output (7)”* [19]. The results of our analysis coincide with the above statement in that: (i) Pearson’s correlation analysis identified a linear correlation between subcutaneous glucose time series and heart rate and systolic blood pressure for the majority of patients, and (ii) the introduction of such variables into the glucose predictive function improved the clinical accuracy of short-term glucose predictions, as it is captured by the CG-EGA analysis.

The selected LSTM model produces highly accurate predictions for both input cases, Case 1 and Case 2, and, although a head-to-head comparison of the different approaches cannot be directly applied since the underlying datasets and the associated clinical studies settings differs substantially, the results presented herein compare well with state-of-the-art performances attained for the same research problem. Nonetheless, we are currently refining the dimensionality reduction step, the prediction function and the training approach targeting the minimisation of the errors in the critical zones of hypoglycaemia and hyperglycaemia, fully exploiting not only linear but also nonlinear correlation between the glucose time series and the investigated herein cardiac indices time series, and, in parallel, investigating adaptive learning algorithms.

Acknowledgement. This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EDK-03990).

References

1. Frayn, K.N.: Metabolic Regulation: A Human Perspective. 3rd edn., pp. 306–308. Wiley-Blackwell, UK (2010)
2. Holt, R.I.G., et al.: The management of type 1 diabetes in adults. A consensus report by the American Diabetes Association (ADA) and the European association for the study of diabetes (EASD). Diab. Care **44**(11), 2589–2625 (2021)
3. American Diabetes Association Professional Practice Committee: 6. Glycemic targets: standards of medical care in diabetes. Diab. Care **45**(Supplement_1), S83–S96 (2022)
4. Amiel, S.A.: The consequences of hypoglycaemia. Diabetologia **64**(5), 963–970 (2021). https://doi.org/10.1007/s00125-020-05366-3
5. Khunti, K., et al.: Hypoglycemia and risk of cardiovascular disease and all-cause mortality in insulin-treated people with type 1 and type 2 diabetes: a cohort study. Diab. Care **38**(2), 316–322 (2014)
6. Tsichlaki, S., et al.: Type 1 diabetes hypoglycemia prediction algorithms: systematic review. JMIR Diab. **7**(3), e34699 (2022)
7. Felizardo, V., et al.: Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction – a systematic literature review. Artif. Intell. Med. **118**, 102–120 (2021)
8. Montaser, E., et al.: Seasonal local models for glucose prediction in type 1 diabetes. IEEE J. Biomed. Health Inform. **24**(7), 2064–2072 (2020)
9. Rabby, M.F., et al.: Stacked LSTM based deep recurrent neural network with Kalman smoothing for blood glucose prediction. BMC Med. Inf. Decis. Making **21**(1), 101 (2021)
10. Schiavon, M., et al.: Modeling subcutaneous absorption of long-acting insulin glargine in type 1 diabetes. IEEE Trans. Biomed. Eng. **67**(2), 624–631 (2020)
11. Muñoz-Organero, M., et al.: Learning carbohydrate digestion and insulin absorption curves using blood glucose level prediction and deep learning models. Sens. (Basel) **21**(14), 4926 (2021)
12. Xie, J.: Benchmarking machine learning algorithms on blood glucose prediction for type I diabetes in comparison with classical time-series models. IEEE Trans. Biomed. Eng. **67**(11), 3101–3124 (2020)
13. Rodríguez-Rodríguez, I., et al.: On the possibility of predicting glycaemia 'on the fly' with constrained IoT devices in type 1 diabetes mellitus patients. Sens. (Basel) **19**(20), 4538 (2019)
14. Vettoretti, M., et al.: Advanced diabetes management using artificial intelligence and continuous glucose monitoring sensors. Sensors **20**(14), 3870 (2020)
15. Yotam, A., et al.: Clinically accurate prediction of glucose levels in patients with type 1 diabetes. Diab. Technol. Therap. **22**, 562–569 (2020)

16. Brownlee, J.: Deep Learning for Time series Forecasting. Machine Learning Mastery (2018)
17. Shou, Y., et al.: Fst and exact warping of time series using adaptive segmental approximations. Mach. Learn. **58**(2), 231–267 (2005)
18. Kovatchev, B.P., et al.: Evaluating the accuracy of continuous glucose-monitoring sensors: continuous glucose–error grid analysis illustrated by TheraSense freestyle navigator data. Diab. Care **27**(8), 1922–1928 (2004)
19. Frier, B.M., et al.: Hypoglycemia and cardiovascular risks. Diab. Care **34**(2), 132–137 (2011)

Less is More: Leveraging Digital Behavioral Markers for Real-Time Identification of Loneliness in Resource-Limited Settings

Md. Sabbir Ahmed(✉) and Nova Ahmed

Design Inclusion and Access Lab (DIAL), North South University, Dhaka, Bangladesh
msg2sabbir@gmail.com, nova.ahmed@northsouth.edu

Abstract. The resource-constrained nature of developing regions and also the positive impact of early intervention show the need for a minimal and faster system to identify loneliness. However, existing pervasive device-based promising systems' requirement to run in the background for prolonged periods can be costly in terms of resources and also may not be effective for early intervention. Thus, we conducted a study ($N = 105$) in Bangladesh by developing a minimal system that can retrieve the past 7 days' app usage behavioral data within a second ($\text{Mean} = 0.31$ s, $\text{SD} = 1.1$ s). Leveraging only the instantly accessed data, we developed models through features selected by 3 different methods and exploration of 14 diverse machine learning (ML) algorithms including 8-tree-based algorithms. We found that the Gaussian Naïve Bayes model, developed by filter method Information Gain selected features, can identify 90.7% of lonely participants correctly with an F1 score of 82.4%. Through SHapley Additive exPlanations (SHAP), we explained the ML models showing how the features impacted the model's outcome. Due to being minimal, faster, and explainable, our system can play a role in resource-limited settings for early identification of loneliness which may create a positive impact by mitigating the loneliness rate.

Keywords: Loneliness · Smartphone · Resource-limited settings · Real-time · Explainable ML

1 Introduction

Mental health of people in Low- and Middle-Income Countries (LMIC) is much neglected than in high-income countries [1]. For example, compared to low-income countries, high-income countries have more than 35 times mental health workers for every 100,000 people [1]. The large divide persists in mental health research also where only 6% of literature emanated from the LMIC [2]; where, over 80% of people having a mental disorder, and also 80% of the world's total population reside in LMIC [3], highlighting the necessity of prioritizing research in that context. Loneliness, a perceived feeling of being separated from others [12], is linked with poor sleep, dementia, depression, and suicidal ideation [5] which may deteriorate if it remains for a prolonged

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 460–476, 2023.
https://doi.org/10.1007/978-3-031-34586-9_31

period. In Bangladesh, around 43% of university students feel high loneliness [7]. However, there are only 270 psychiatrists and 565 psychologists in the country having over 160 million people [6] and 1.2 million students of universities [34]. For the early identification of loneliness in resource-limited settings like Bangladesh, the smartphone can be incorporated due to its high availability among the youth where 86.62% of university students in Bangladesh own smartphones [14].

Given the usability of the digital behavioral markers, researchers explored pervasive devices for a wide range of problems including loneliness [8–11, 21]. Existing studies leveraged phone usage [8–10, 21], smartphone sensed [9, 11], Fitbit sensed [9] and other systems [21] retrieved data to develop machine learning (ML) models for loneliness identification. While Austin et al. [21] focused on older adults, other studies [8–11] used youth as the samples who have higher loneliness [22]. ML models of the existing studies show promising performance. For example, Doryab et al. [9] found an accuracy of over 80% in identifying lonely participants. The positive impact of early intervention of psychological problems [20] shows the need for a faster system. However, the main limitation of the existing pervasive device-based system is the requirement for a prolonged data collection period (e.g., 2 weeks [10], 10 weeks [8], 16 weeks [9], and several months [21]) which may not be effective for early intervention. Also, existing systems' [8–11, 21] requirement for running a tool in the background throughout the whole study period may introduce research reactivity problems (e.g., Hawthorne effect). In addition, due to the consumption of much battery power of the background services [23], there may be significant barriers to having quality data in low-resource settings where electricity and internet services are limited [25].

To overcome these limitations, we present a minimal and unobtrusive system that can identify loneliness in real-time. Our study makes 3-fold contributions:

- Leveraging our tool's instantly (Mean = 0.31 s, SD = 1.1 s) retrieved app usage behavioral markers only, our ML model can identify 90.7% of lonely participants correctly (F1 = 82.4%). As far as we know, compared to any other existing pervasive device-based systems, to identify lonely students, our system is faster and minimal which can enable it to play a significant role in low-resource settings.
- We developed ML models by 14 classification algorithms including the linear, non-linear, Stacking and Weighted Voting algorithms. We selected important features by 1 feature selection (FS) algorithm from each of the 3 main FS methods [19]. With comprehensive exploration, we presented a parsimonious ML model developed with around 6 features (Mean = 5.8, SD = 1.3) which has a sensitivity and specificity of over 70%. Due to having a lower number of features, this finding can be useful to have a more resource-insensitive system.
- Through SHapley Additive exPlanations (SHAP), we present how different behavioral markers impacted the predicted class. We discuss the findings of explainable ML which can facilitate mental healthcare professionals to understand lonely students more and take steps in intervention accordingly.

2 Related Work

2.1 Relation of App Usage Behavioral Markers with Loneliness

Smartphone usage behavior has a relation with loneliness [9]. In a smartphone, there remains a diverse set of apps and each app's unique features keep it in a distinct category [27]. Apps such as Facebook, Instagram, and Snapchat are in the Social Media category and their higher usage has a relation to lower loneliness [26]. However, depending on the category, there is a variation in users' behavior [27] and also the relation between loneliness and app usage [8, 10, 28]. For instance, while the number of messages has a negative relation, browsing searchers at late night has a positive relation with loneliness [10]. Even within the same category, there can be variation in relation depending on the behavioral markers. For instance, though loneliness has a negative association with the number of incoming calls, the number of missed calls does not have any significant association [10]. This reflects the importance of leveraging different behavioral markers while developing computational models to identify loneliness.

2.2 Identifying Loneliness Through Pervasive Devices

Comparatively, a significantly higher number of research about the identification of particular mental problem through sensing devices has focused on depression as presented by a recent systematic review [30]. However, loneliness has an effect on depression [31] and depression can be mitigated by mitigating loneliness [32] which presents the importance of prioritizing research on loneliness identification.

Nevertheless, little research has been conducted in identifying loneliness through unobtrusive ways. Some of these studies [8, 11] leveraged smartphone data. In a study [8] on 46 participants, researchers correctly identified 67.89% of the lonely participants by their smartphone usage and sensed data. Another study [10] used 2 weeks' smartphone usage, WIFI, and Bluetooth sensed data and their model correctly identified loneliness with an accuracy of 90%. However, due to having only 9 participants and due to unavailable details of their ML model (e.g., building and evaluation details of the classifier) [10], their findings may not be generalizable. In a previous study [11], utilizing the smartphone sensed geospatial data, an ML model distinguished the lonely from the non-lonely with an AUC value of .74. But their study has some limitations. For instance, the researchers [11] used the response of a single question as ground truth to group the lonely and non-lonely which may not be a standard method such as the validated UCLA Loneliness Scale-8 (ULS-8) [12]. In addition, to get stable performance, they needed several days' data which was collected by running their tool in the background for an equal number of days. In other previous studies also, researchers needed to use the data for several weeks (e.g., 10 weeks [8], 16 weeks [9]) running the tool for the whole study period. However, loneliness is associated with physical and psychological problems [5] where early identification and intervention can play a role in mitigating the severe consequences as early intervention has a positive impact [20].

3 Behavioral Data Collection Tool

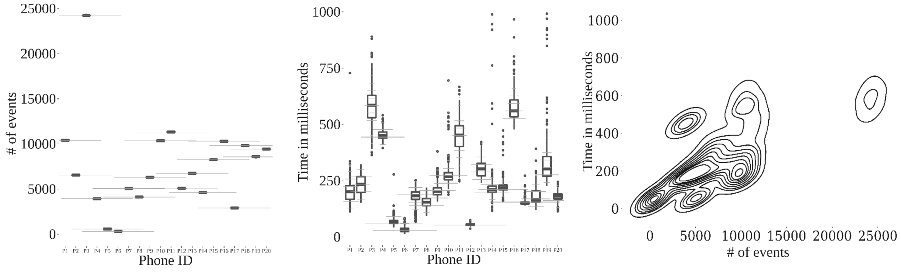
3.1 Development and Validation of the Tool

To retrieve the actual app usage behavioral markers (e.g., launch) unobtrusively, we developed an Android app. We chose the Android platform as it is used by 95.68% [41] phone users of Bangladesh. To retrieve the foreground and background events' data, we used several functions of the Java Class *UsageStatsManager* [33]. However, since app usage events are kept only for a few days [33], our tool can retrieve instantly (Sect. 3.2.) the app usage data of the past 7 days.

There were 3 steps in the app testing phase. We compared our tool's retrieved app usage data with the manually calculated data and retrieved data of such tools (e.g., [13]) available in Google Play Store which needs to run in the background. Considering the variations of phones, we also checked our app's retrieved data in 9 different phones.

3.2 Required Time to Retrieve Data

To estimate the generalizable time required to retrieve the foreground and background events from smartphones, we tested our tool on 20 smartphones of 19 different models and 8 different operating system (OS) versions of Android. From each phone, we collected the past 7 days' app usage data 500 times, and in total, we calculated the required time 10,000 times. The average number of retrieved foreground and background events was 7,447.61 (Min. = 306, Max. = 24,297, Median = 6,641, SD = 4986.62) (Fig. 1(a)) and on average, it took 307.94 ms (ms) (Min. = 13 ms, Max. = 61,087 ms, Median = 211 ms, SD = 1103.91 ms) (Fig. 1(b)).



(a) Number of retrieved events (b) Required time (c) KDE plot showing relation between time and number of events

Fig. 1. Performance of the data collection tool. KDE: Kernel Density Estimation. In figure b and figure c, 97 instances where the required time was more than 1000 ms were truncated to make the smaller values visible and we did not scale the required time to present the actual data.

To understand the factors that can impact the time in data retrieval, we explored the correlation of required time with 20 phones' API level and the number of retrieved events. We calculated the Spearman correlation coefficient (r_s) as the data did not satisfy the assumptions of the parametric test. We found no significant relation between the Android API level and required time ($r_s = .18$, $p = .44$). In exploring the relation with the number of events, we used the average number of events and the required average time for each phone as within each phone, there was almost no variation in the number of retrieved events (Fig. 1(a)). We found that the events' number has a significant positive relation with the required time to retrieve data ($r_s = .56$, $p = .0096$) (Fig. 1(c)). After that, to estimate the plausible number of events on the students' phones, we used our constructed dataset for this study having 105 students (please, see Sect. 4.1 and Sect. 4.3 for details). In the dataset, on average, there were 8,174.04 events (SD = 4972.5). Among our data retrieval for 10,000 times, there were 4,500 instances for which the number of retrieved events was more than 8,000, and to retrieve this large number of events, our app needed 430.31 ms (SD = 1596.455 ms) which reveals that on average, our app can retrieve past 7 days' app usage data within a second.

4 Methodology

4.1 Research Ethics and Participants

Our study was approved by the Center for Research and Development of a university. All participants provided their consent before voluntary participation and data were stored in secure storage where only the researchers of this project can access. We collected data during the COVID-19 pandemic from January to June 2021. From 8 educational institutes, 105 students participated whose mean age was 22.3 years (SD = 1.57) and they were from 33 districts among the 64 districts of Bangladesh.

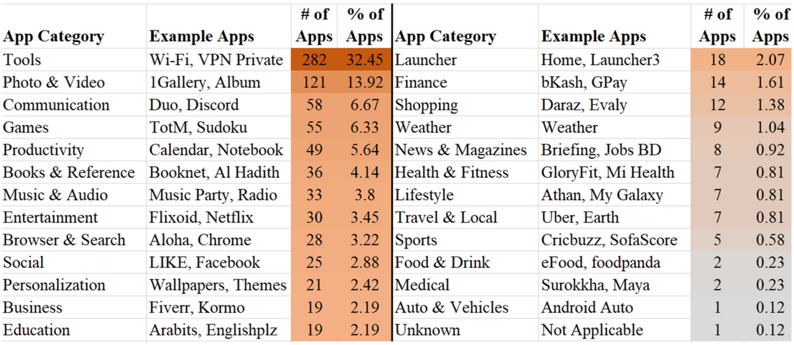
4.2 Categorization of Lonely and Non-lonely Participants

To understand loneliness and to use as the ground truth labels for ML models, we used the score of ULS-8 [12] which has been used for identification of loneliness in different countries (e.g., USA [12], Bangladesh [15]) showing the validity. There are 8 items and participants responded to the items through our developed app while donating app usage data. The options to respond for each item is Never (score 1), Rarely (score 2), Sometimes (score 3), and Always (score 4). Having a score of more than 16 means there is at least one loneliness measuring item that was bothered sometimes. Following the previous studies [9, 15], we categorized the participants having ULS-8 score of more than 16 into the lonely, and others were kept in the non-lonely group.

4.3 Feature Extraction and App Usage Behavioral Markers

In 7 days, 105 students used 867 apps and there were 868,636 foreground and background events' data from which we extracted the features.

App categories. For app categorization, at first, we retrieved the developers' preferred category available in Play Store using a Java HTML parser. Students used several apps which were not available there, and thus, using the package name, we explored those apps' features in online app stores (e.g., apkmonk.com) and developers' websites. Finally, we did categorization by understanding the process of previous studies (e.g., [17]) and through a discussion with 2 graduate students of engineering faculty. We found 105 students



App Category	Example Apps	# of Apps	% of Apps	App Category	Example Apps	# of Apps	% of Apps
Tools	Wi-Fi, VPN Private	282	32.45	Launcher	Home, Launcher3	18	2.07
Photo & Video	1Gallery, Album	121	13.92	Finance	bKash, GPay	14	1.61
Communication	Duo, Discord	58	6.67	Shopping	Daraz, Evaly	12	1.38
Games	TotM, Sudoku	55	6.33	Weather	Weather	9	1.04
Productivity	Calendar, Notebook	49	5.64	News & Magazines	Briefing, Jobs BD	8	0.92
Books & Reference	Booknet, Al Hadith	36	4.14	Health & Fitness	GloryFit, Mi Health	7	0.81
Music & Audio	Music Party, Radio	33	3.8	Lifestyle	Athan, My Galaxy	7	0.81
Entertainment	Flixoid, Netflix	30	3.45	Travel & Local	Uber, Earth	7	0.81
Browser & Search	Aloha, Chrome	28	3.22	Sports	Cricbuzz, SofaScore	5	0.58
Social	LIKE, Facebook	25	2.88	Food & Drink	eFood, foodpanda	2	0.23
Personalization	Wallpapers, Themes	21	2.42	Medical	Surokkha, Maya	2	0.23
Business	Fiverr, Kormo	19	2.19	Auto & Vehicles	Android Auto	1	0.12
Education	Arabits, Englishplz	19	2.19	Unknown	Not Applicable	1	0.12

Fig. 2. Example apps along with the number (#) and percentage (%) of apps of each category.

using 867 apps of 26 categories (Fig. 2). Most apps were from Tools (32.45%) and the least apps were from the Auto & Vehicles (0.12%) category (Fig. 2).

Ratio of hamming distance. As uniqueness in terms of app usage varies among smartphone users (e.g., between the depressed and non-depressed [17]), in the case of each student, we calculated the ratio of the hamming distance from the nearest lonely to the nearest non-lonely student. To get an unbiased ML model, we did not consider the group (e.g., non-lonely) of that student while calculating the distance: $DL_{ij} = (A_i \cup A_j) - (A_i \cap A_j)$ where DL_{ij} denotes the distance of the i^{th} student from the j^{th} student of the lonely group; A_i and A_j denote the set of apps used by the i^{th} and j^{th} students respectively. In this way, we calculated the minimum distance DL_i of the i^{th} student in the lonely group. We followed the same process to calculate the minimum distance DNL_i of the i^{th} student in the non-lonely group. Finally, we calculated the ratio of hamming distance for the student: $\frac{DL_i}{DNL_i}$. The main motivation behind using ratio, instead of global distance (minimum distance among all participants regardless group) is that it tells us how much or less a participant is unique compared to the lonely and non-lonely participants which are intuitively more informative.

Other behavioral data: For total smartphone usage (i.e., regardless of app category) and each of the 26 app categories, we also extracted duration, frequency of launch, duration per launch, launch per app, duration per app, and session data to extract the participants' app usage patterns. In addition, as app usage behavior varies by days [17], we extracted features regarding the difference in app usage between weekdays and weekends dividing the days (Weekdays: Sunday to Thursday, Weekends: Friday and Saturday) based on the working schedule of Bangladesh.

Characteristics of the features: Apart from the features using the 24 h data, for each of the above-mentioned behavioral markers, we also extracted diurnal usage data dividing days into 4 equal time periods: Night: 00:01–6:00; Morning: 6:01–12:00; Afternoon: 12:01–18:00; Evening: 18:01–00:00. After that, we calculated 8 different data from the diurnal usage to be used as features: minimum, maximum, range, mean, standard deviation, entropy, skewness, and kurtosis. To understand the app usage patterns more, we extracted features in two ways using the entropy formula. Firstly, for the diurnal usage data, we calculated the entropy $E_t(j) = -\sum_{t=1}^{4} P_d(i) log P_d(i)$ where for the student j who has an unequal spending duration in each of the 4 time periods, E_t will be lower compared to the student who has an equal spending duration in each time period. Let's say, a student has 1, 1, 3, and 3 min whereas another student has 2, 2, 2, and 2 min per app spending duration on apps of a category in the night, morning, afternoon, and evening periods respectively. Then, the 1^{st} student will have an entropy E_t of 1.81 which is lower than the entropy E_t of 2.0 of the 2^{nd} student presenting that the 1^{st} student is more focused on the phone during certain time periods and also has variation in app usage over the day compared to the 2^{nd} student. Secondly, we calculated entropy $E_{dl}(j)$ for the student j on the basis of spending duration and frequency of launching of each app of an app category.

$E_{dl}(j) = -\frac{1}{2}(\sum_{i=1}^{n} P_d(i) log P_d(i) + \sum_{i=1}^{n} P_l(i) log P_l(i))$; here, P_d and P_l denote the probability to use i^{th} app based on spending duration and launch respectively.

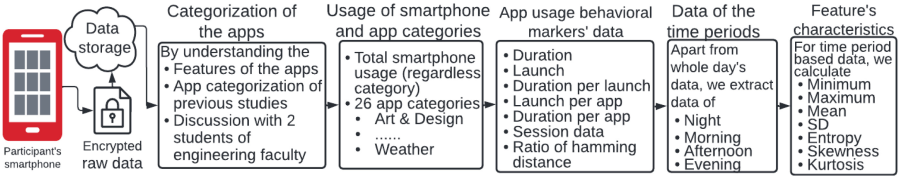
4.4 Feature Selection

As there is no feature selection (FS) method that can find the best set of features ensuring the maximum performance of the ML models, we explored 1 FS approach from each of the 3 main FS categories [19]: wrapper, filter, and embedded method. As a wrapper method, we used the Boruta where, unlike the minimal-optimal methods, all-relevant features are selected [18]. In Boruta, Random Forest (RF) is used as the base estimator [18] and it is suggested to use 3 to 7 as the base estimator's maximum depth [35]. As the filter and embedded methods, we used the Information Gain (IG) and RF respectively. However, unlike Boruta, these two methods do not inform a fixed set of features that can have the best performance. Hence, we set the lower boundary of features using the 1 in 10 rule [38] where 5 features are to be selected due to having 54 lonely participants in our study. We increased the number of features gradually up to 20 and did not increase further to prevent the possibility of having overfitted models.

4.5 Model Development, Validation, and Explanations

We preferred machine learning (ML) to deep learning since ML models have higher transparency. We extracted the features (Fig. 3(a)) ourselves and also explained the ML models which can be insightful, particularly for mental healthcare professionals. As there is no ML model which can fit for all solutions, we developed models by a set of classification algorithms where both linear and non-linear including 8-tree based algorithms were explored: AdaBoost, CatBoost, Decision Trees, Extra Tree, Extreme Gradient Boosting (XGBoost), Gaussian Naïve Bayes (NB), Gradient Boosting (GB), K-Nearest Neighbor (KNN), Light GBM, Logistic Regression (Logit), RF, C-Support Vector Classifier (SVC). In addition, as a baseline, we used a Dummy classifier.

To develop the models, we used the nested cross-validation approach which shows a generalizable and unbiased performance [16]. In the outer loop, we used the Leave-One-Out-Cross-Validation (LOOCV) method which has a lower variance [36] and where we divided the dataset into n equal portions presenting each participant's data. In the inner loop, $n - 1$ participants' data were used for 2 purposes: to select a set of important


Participant's smartphone
Data storage
Encrypted raw data
Categorization of the apps
By understanding the
• Features of the apps
• App categorization of previous studies
• Discussion with 2 students of engineering faculty
Usage of smartphone and app categories
• Total smartphone usage (regardless category)
• 26 app categories
• Art & Design
• ......
• Weather
App usage behavioral markers' data
• Duration
• Launch
• Duration per launch
• Launch per app
• Duration per app
• Session data
• Ratio of hamming distance
Data of the time periods
Apart from whole day's data, we extract data of
• Night
• Morning
• Afternoon
• Evening
Feature's characteristics
For time period based data, we calculate
• Minimum
• Maximum
• Mean
• SD
• Entropy
• Skewness
• Kurtosis

(a) Behavioral markers and the process for extraction of the features

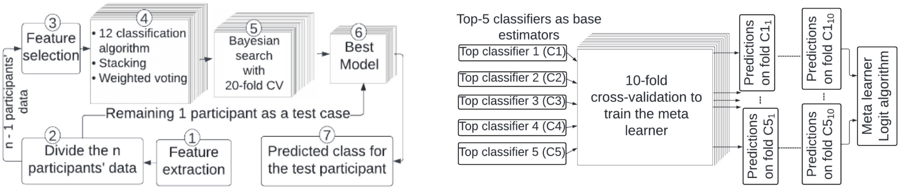

③ Feature selection
④ • 12 classification algorithm • Stacking • Weighted voting
⑤ Bayesian search with 20-fold CV
⑥ Best Model
n - 1 participants' data
Remaining 1 participant as a test case
② Divide the n participants' data
① Feature extraction
⑦ Predicted class for the test participant
Top-5 classifiers as base estimators
Top classifier 1 (C1)
Top classifier 2 (C2)
Top classifier 3 (C3)
Top classifier 4 (C4)
Top classifier 5 (C5)
10-fold cross-validation to train the meta learner
Predictions on fold C1₁
Predictions on fold C5₁
Predictions on fold C1₁₀
Predictions on fold C5₁₀
Meta learner Logit algorithm

Fig. 3. Overview of the ML model development process.

features and to tune the hyper-parameters by the Bayesian search optimization technique using 20-fold CV instead of the LOOCV method to reduce the time complexity. In the Bayesian optimization technique, the informed search technique is used where the next step is taken based on the performance of the previous step and unlike Grid Search, this method does not need to explore every combination which makes it faster. After finding the best estimator, we predict the remaining 1 participant's class who was not included in FS and hyper-parameter tuning steps (Fig. 3(b)). To predict each of the 105 student's class, we repeat the same process. It can be noted that to build the ML models, we used open-sourced Python libraries.

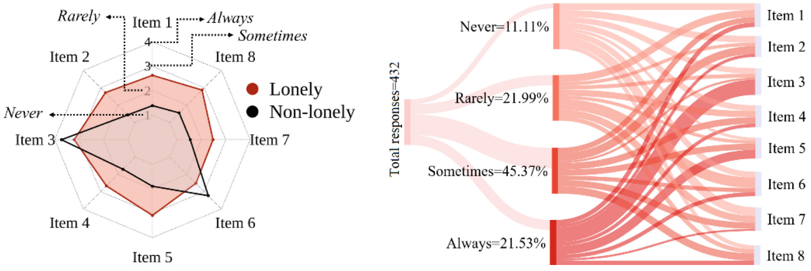
After developing the individual ML models based on the aforementioned 12 classification algorithms, using the best 5 classification algorithms, we developed an ensemble model Stacking and Weighted Voting. In the Stacking classifier, the predictions of the individual estimators were stacked to train the meta-learner Logit (Fig. 3(c)). On the other hand, in the Weighted Voting classifier, we calculated the weights by dividing the training data into 10-folds. The weight was multiplied by the final predicted probability of each of the top-5 classifiers for the test participant. After that, the final class for the participant is decided based on the soft-voting.

We evaluated the models' performance by comparing the models' predicted class with the ground truth class based on ULS-8 score. We calculated sensitivity and specificity which present how many of the lonely and non-lonely were identified correctly respectively. In addition, we presented precision which informs us the percentage of predicted lonely participants was truly lonely. We focused on maximizing the F1 score which is the harmonic mean of precision and sensitivity. To explain the ML models, we used the SHAP [4] which works based on the concept of cooperative game theory.

5 Findings

5.1 Participants' Loneliness

Among 105 participants, 54 participants (51.4%) were lonely and 51 participants (48.6%) were non-lonely (please, see Sect. 4.2 for categorization process). Except for the reverse items (3rd item: *I am an outgoing person*; 6th item: *I can find companionship when I want it*), lonely participants had a much higher frequency of bothering with loneliness (Fig. 4(a)). For instance, lonely participants' average frequency of feeling isolated from others (5th item) was more than sometimes which was rarely in the case of non-lonely participants (Fig. 4(a)). Also, lonely participants' most (45.37%) responses for the items were sometimes and 21.53% responses were always (Fig. 4(b)) presenting around 67% responses containing feeling of loneliness at least sometimes.



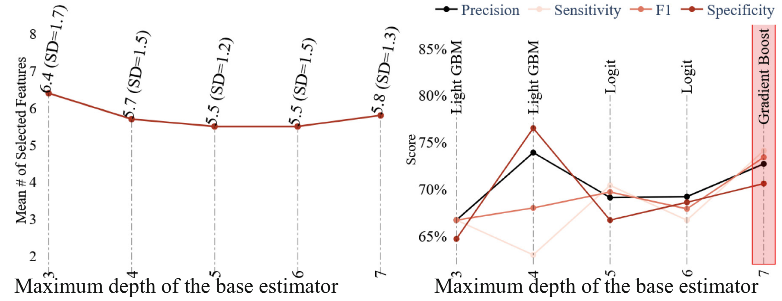
(a) (b)

Fig. 4. (a) The difference between the lonely and non-lonely students in frequency of the ULS-8 scale's items' appearance. (b) The link between the items and the frequency of appearance, based on the 432 (54 lonely students * responses of the 8 items) responses of the lonely students.

5.2 ML Models' Performance

Exploring the Boruta selected features, we found that the mean number of selected important features varied with the variation of the maximum depth of the base estimator Random Forest (RF) algorithm. In each depth, the average number of selected features in each iteration of LOOCV was around 6 (Fig. 5(a)). However, the performance of the models varied largely where we found the minimum F1 score of 66.7% at depth 3 and a maximum of 73.4% at depth 7 (Fig. 5(b)). At depth 7, the average number of selected features in LOOCV was 5.8 (SD = 1.3) and the best performing model among the explored classification algorithms was Gradient Boosting (GB) which had sensitivity, specificity, and precision of 74.1%, 70.6%, and 72.7% respectively (Fig. 5(b)). These present that the GB model identified 74.1% lonely and 70.6% non-lonely correctly. Also, the predicted lonely class was correct in 72.7% of cases.

In the models based on the filter method Information Gain (IG) selected 5 to 11 important features, the best models' sensitivity score varied from 53.7% to 68.5%, and the specificity score varied from 56.9% to 70.6% (Fig. 6(a)). At the number of features



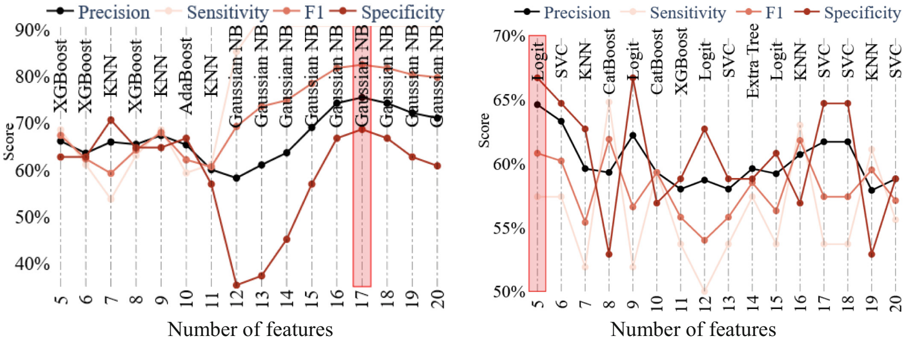
(a) Mean number of selected important features in each iteration of LOOCV with varying depth

(b) Best model's performance with varying depth

Fig. 5. (a) Wrapper method Boruta selected important features and (b) performance of the best models. The rectangle box presents the model which had optimal performance.

12, though the specificity (35.3%) reached a minimum, sensitivity (85.2%) increased. Gradually, increasing the number of selected important features, we found the best performance when the Gaussian Naïve Bayes (NB) was built on 17 important features selected in each iteration of LOOCV (Fig. 6(a)). The NB model identified 90.7% of lonely students correctly (sensitivity = 90.7%) and also predicted lonely students were truly lonely in 75.4% of cases (precision = 75.4%).

On the other hand, in the embedded method RF selected 17 features, the C-Support Vector Classifier (SVC) model performed best which had a sensitivity and specificity of 53.7% and 64.7% respectively (Fig. 6(b)). Among the RF selected important features from 5 to 20 (range setting process is available in Sect. 4.4), we found the best model while the Logit model was developed based on 5 features. The Logit model had a sensitivity, specificity, and precision score of 57.4%, 66.7%, and 64.6% respectively.



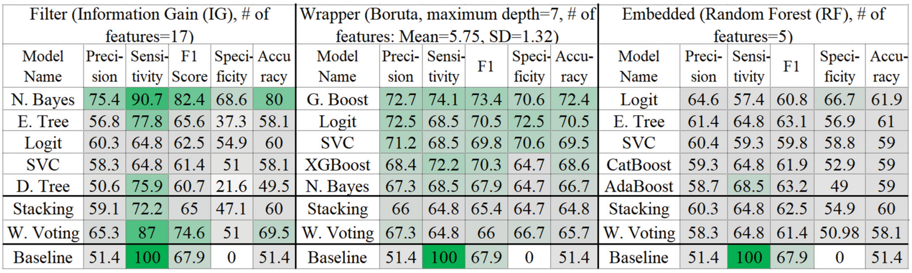
(a) Filter method Information Gain (IG)

(b) Embedded method Random Forest (RF)

Fig. 6. Best models' performance in a varying number of selected important features by the (a) IG and (b) RF FS methods. The rectangle box presents the model which had optimal performance.

Interestingly, it appears that the Boruta selected important features-based ML models have a higher performance with a lower number of features. For instance, with 6 features selected by the IG, the XGBoost performed best having a sensitivity of 61.1% (Fig. 6(a)) while the RF selected 6 features based best performing model SVC had a sensitivity of 57.4% (Fig. 6(b)). However, in all of the explored maximum depths of the Boruta, the number of selected features was around 6 (Fig. 5(a)) where the minimum and maximum sensitivity were 63% and 74.1% respectively and the maximum performing model was GB (Fig. 5(b)) as aforementioned. This presents GB as a parsimonious model having higher predictability with a lower number of features.

Among all models of all feature selection (FS) methods, we found the best performance in terms of F1 score, sensitivity, precision, and accuracy from NB model (sensitivity = 90.7%, F1 score = 82.4%) which was developed based on 17 important features selected by the IG (Fig. 7). However, the specificity of the model was 68.6% whereas the GB model based on the Boruta selected features had a specificity score of 70.6% (Fig. 7) presenting relatively higher ability to identify the non-lonely participants.



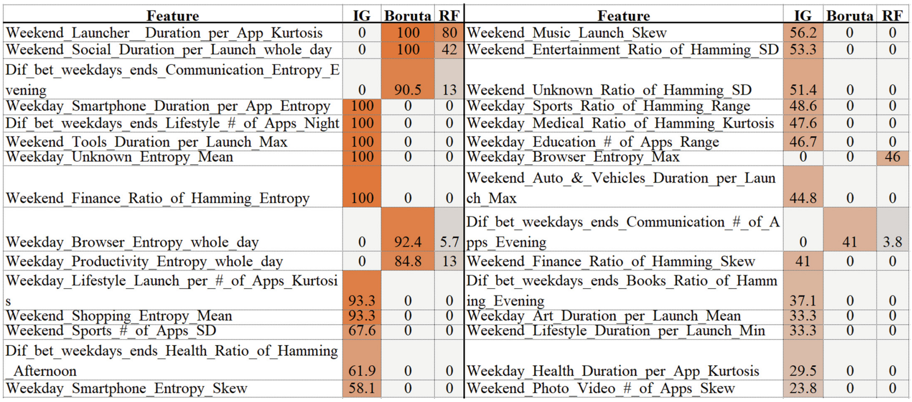
Filter (Information Gain (IG), # of features=17)						Wrapper (Boruta, maximum depth=7, # of features: Mean=5.75, SD=1.32)						Embedded (Random Forest (RF), # of features=5)					
Model Name	Precision	Sensitivity	F1 Score	Specificity	Accuracy	Model Name	Precision	Sensitivity	F1	Specificity	Accuracy	Model Name	Precision	Sensitivity	F1	Specificity	Accuracy
N. Bayes	75.4	90.7	82.4	68.6	80	G. Boost	72.7	74.1	73.4	70.6	72.4	Logit	64.6	57.4	60.8	66.7	61.9
E. Tree	56.8	77.8	65.6	37.3	58.1	Logit	72.5	68.5	70.5	72.5	70.5	E. Tree	61.4	64.8	63.1	56.9	61
Logit	60.3	64.8	62.5	54.9	60	SVC	71.2	68.5	69.8	70.6	69.5	SVC	60.4	59.3	59.8	58.8	59
SVC	58.3	64.8	61.4	51	58.1	XGBoost	68.4	72.2	70.3	64.7	68.6	CatBoost	59.3	64.8	61.9	52.9	59
D. Tree	50.6	75.9	60.7	21.6	49.5	N. Bayes	67.3	68.5	67.9	64.7	66.7	AdaBoost	58.7	68.5	63.2	49	59
Stacking	59.1	72.2	65	47.1	60	Stacking	66	64.8	65.4	64.7	64.8	Stacking	60.3	64.8	62.5	54.9	60
W. Voting	65.3	87	74.6	51	69.5	W. Voting	67.3	64.8	66	66.7	65.7	W. Voting	58.3	64.8	61.4	50.98	58.1
Baseline	51.4	100	67.9	0	51.4	Baseline	51.4	100	67.9	0	51.4	Baseline	51.4	100	67.9	0	51.4

Fig. 7. Performance of the top-5 classifiers, based on the best (in terms of ML models' performance) set of features in each FS method. "# of features" present the number of features used in each iteration of LOOCV. D: Decision, E: Extra, G: Gradient, N: Naïve, W: Weighted.

Based on each FS method's top-5 classifiers, we developed Stacking and Weighted Voting models. Among these, IG selected features based Weighted Voting model had a higher performance which identified 87% lonely and 51% non-lonely students correctly (Fig. 7) having a balanced accuracy ($\frac{Sensitivity+Specificity}{2}$) of 69%. Though all the models had a higher performance than the baseline Dummy classifier's balanced accuracy 50%, and specificity of 0%, the Stacking, and Weighted Voting classifiers' performance were not higher than the best classifier of each FS method (Fig. 7).

5.3 Explanation of the ML Models

Exploring the top-30 important features which were used for the ML model development, we did not find a common feature that appeared in each of the 3 FS methods (Fig. 8). This is reflected in the performance of the ML models also where we found a higher variation across the FS methods (Fig. 7) which presents FS methods' different mechanisms of selecting the important features and also explains the rationale behind the exploration of 3 different FS methods. Among the top-30 features, 3 (10%) features were regarding the whole day whereas the remaining 27 (90%) important features were based on the four time periods of a day (night, morning, afternoon, and evening) (Fig. 8). This presents that the diurnal usage data contains more information in identifying the differences between the lonely and non-lonely students. Similarly, compared to the features on total smartphone usage regardless of the app category (6.67%), there were a higher number of important features in the case of the app categories (93.3%).



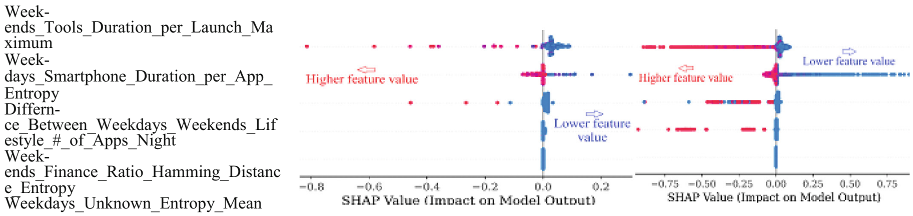
Feature	IG	Boruta	RF	Feature	IG	Boruta	RF
Weekend_Launcher__Duration_per_App_Kurtosis	0	100	80	Weekend_Music_Launch_Skew	56.2	0	0
Weekend_Social_Duration_per_Launch_whole_day	0	100	42	Weekend_Entertainment_Ratio_of_Hamming_SD	53.3	0	0
Dif_bet_weekdays_ends_Communication_Entropy_Evening	0	90.5	13	Weekend_Unknown_Ratio_of_Hamming_SD	51.4	0	0
Weekday_Smartphone_Duration_per_App_Entropy	100	0	0	Weekday_Sports_Ratio_of_Hamming_Range	48.6	0	0
Dif_bet_weekdays_ends_Lifestyle_#_of_Apps_Night	100	0	0	Weekday_Medical_Ratio_of_Hamming_Kurtosis	47.6	0	0
Weekend_Tools_Duration_per_Launch_Max	100	0	0	Weekday_Education_#_of_Apps_Range	46.7	0	0
Weekday_Unknown_Entropy_Mean	100	0	0	Weekday_Browser_Entropy_Max	0	0	46
Weekend_Finance_Ratio_of_Hamming_Entropy	100	0	0	Weekend_Auto_&_Vehicles_Duration_per_Launch_Max	44.8	0	0
Weekday_Browser_Entropy_whole_day	0	92.4	5.7	Dif_bet_weekdays_ends_Communication_#_of_Apps_Evening	0	41	3.8
Weekday_Productivity_Entropy_whole_day	0	84.8	13	Weekend_Finance_Ratio_of_Hamming_Skew	41	0	0
Weekday_Lifestyle_Launch_per_#_of_Apps_Kurtosis	93.3	0	0	Dif_bet_weekdays_ends_Books_Ratio_of_Hamming_Evening	37.1	0	0
Weekend_Shopping_Entropy_Mean	93.3	0	0	Weekday_Art_Duration_per_Launch_Mean	33.3	0	0
Weekend_Sports_#_of_Apps_SD	67.6	0	0	Weekend_Lifestyle_Duration_per_Launch_Min	33.3	0	0
Dif_bet_weekdays_ends_Health_Ratio_of_Hamming_Afternoon	61.9	0	0	Weekday_Health_Duration_per_App_Kurtosis	29.5	0	0
Weekday_Smartphone_Entropy_Skew	58.1	0	0	Weekend_Photo_Video_#_of_Apps_Skew	23.8	0	0

Fig. 8. Top-30 (average presence in each FS method) important features among the features used for the top-5 classifiers of each FS method. The values present the percentage of iterations a feature appeared as important among all iterations of LOOCV. Dif_bet_weekdays_ends: Difference between weekends and weekends, SD: Standard Deviation.

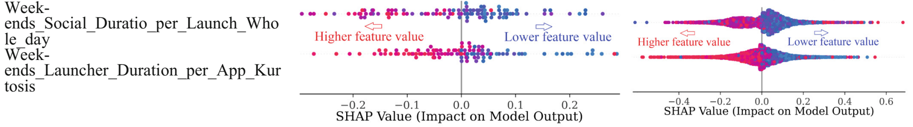
Exploring the features' data characteristics, we found more than half of the features (53.3%) among the top-30 contain the ratio of hamming distance and entropy data (Fig. 8) presenting these features' higher ability to differentiate the lonely and non-lonely students through the ML models. On the other hand, we found almost the same number of solely weekdays (12 features, 40%) and weekends' (13 features, 43.33%) data-based features (Fig. 8) which presents equal importance of weekdays and weekends' data.

Summary plot of the SHAP analysis showed that on the weekend when the Tools app category's (e.g., *Assistant*) maximum duration per launch among the usage in 4 time periods became lower, students were more likely to be in the lonely group (Fig. 9 (a, b)). In addition, when the entropy regarding duration per app became lower on weekdays, the students were also likely to be lonely. Having a lower entropy means there is a higher variation in spending duration per app over the 4 time periods of weekdays (a detailed discussion on entropy is available in Sect. 4.3).

Apart from these, the SHAP analysis demonstrated that when the difference in the number of used Lifestyle apps (e.g., *Athan*) in night time periods of weekdays and weekends became lower, students were more likely to be in the lonely group (Fig. 9(a, b)) which indicates that over the whole week's night, lonely students were likely to use Lifestyle category's apps. We found that among the students' used 7 apps of this category, 4 apps *(Athan, App Of Ramadan, Prayer Time Quran Qibla Dua Tasbih, Muslim Pro)* were related to prayer. Our findings also showed that the lonely students were more likely to have a lower duration per launch of Social Media apps (e.g., *Facebook*) on the weekend (Fig. 9(c, d)). In addition, lonely students' kurtosis values regarding the



(a) Gaussian Naïve Bayes (NB) model on the test data. (b) Gaussian Naïve Bayes (NB) model on the training data.



(c) Gradient Boosting (GB) model on the test data. (d) Gradient Boosting (GB) model on the training data.

Fig. 9. Shapley summary plot for the NB model (a, b) based on 17 features selected by the IG and for the GB model (c, d) based on the Boruta selected features when the maximum depth of the base estimator was 7. The plot shows the impact of the features (which appeared in all iterations of LOOCV) on the ML models' outcome. In figures b and d presenting models' outcome based on the training data, there is a higher number of feature values since for each iteration of LOOCV, $n - 1$ participants' data were in the training.

duration per Launcher (e.g., *Launcher3*) apps' usage on the weekends in the 4 time periods were lower which presents a lower tail in the distribution of their data.

6 Discussion

Our findings present that it is possible to identify loneliness unobtrusively and accurately (Sensitivity = 90.7%, F1 = 82.4%) within a second (Mean = 0.31 s, SD = 1.1 s). On the other hand, the existing robust systems have the need to run in the background for a prolonged period (Table 1) such as 10 weeks [8] and 16 weeks [9] which may not work for early intervention. As a consequence, lonely students' situations can deteriorate. Moreover, current systems' need of running in the background [8–11, 21], and also the need for access to the sensors such as GPS [8, 9, 11] which consumes significant battery power [23] may make the systems infeasible for the resource-limited settings, especially, where limited electricity and internet are two of the barriers to use technology [25]. However, our system does not need to run in the background. In addition, it relies solely on the instantly accessed computationally cheaper app usage data where simple mathematical formulas (please, see Sect. 4.3) are used to extract the features and does not have any dependency on additional computational models (e.g., usage of conversation detection classifiers to extract conversation related features as used in explored dataset of [8]) for feature extraction. All these make our system minimal which can have potential implications for loneliness identification in resource-limited settings such as in developing and underdeveloped countries.

We explored 14 different classification algorithms and the models were built by the features selected by a FS method from each of the 3 main categories [19]. We found

Table 1. Comparison of our system with existing pervasive device-based systems in classifying the lonely and non-lonely. Existing systems' "data retrieval time" is mentioned based on the systems' description available in the research article. NA: Not Available.

Reference	Sample size (N)	Example of the explored data	System's need to run in background	Duration of the collected data	Data retrieval time	Sensitivity	Accuracy	F1
[8]	46	App usage, location, conversation	Yes	10 weeks	10 weeks	67.89	68.67	66.54
[9]	160	Screen, sleep, steps, location	Yes	16 weeks	16 weeks	80.1	80.2	80.1
[10]	9	App usage, Bluetooth, Wi-Fi sensed	Yes	2 weeks	2 weeks	NA	98.0	NA
Our system	**105**	**Only app usage data**	**No**	**1 week**	**Mean: 307.94 ms**	**90.7**	**80**	**82.4**

that the GB model developed by Boruta selected around 6 features (Mean = 5.8, SD = 1.3) have maximum sensitivity and specificity of 74.1% and 70.6% where the sensitivity is more than 10% compared to the filter method Information Gain (IG) and embedded method RF selected 6 features-based models. One of the plausible reasons for having a higher performance is that the Boruta works by selecting all-relevant features to the target variable [18], unlike the minimal-optimal method which is followed by the filter and embedded methods. Due to having a higher performance with a lower number of features, the GB model appears as a parsimonious model. With the lower features, the requirement of computational resources to develop models decreases [40] and thus, the presented GB model can be used to develop a more resource-insensitive system.

SHAP analysis [4] on our best ML model NB based on the IG selected 17 features showed that the lonely students were more likely to have a higher variation in duration per app on weekdays over the 4 time periods. This finding aligns with the studies conducted through conventional statistical methods showing depressed students' variation of diurnal app usage patterns [17]. The variation can reflect students' mood swings while going through a negative experience [24]. In addition, with the variation of time periods, people stay at different places which is related with the variation in usage behavior of app categories [39]. That being said, aggregated data of the whole day may not contain such contextual information and as a result, that may not be informative enough to find subtle differences between the lonely and non-lonely students. This phenomenon is reflected in important feature analysis where compared to the whole day, we found a much higher number of diurnal usage data-based features of the app categories as important. However, the diurnal usage data of the app categories were unexplored in the previous studies [8–10] to develop models to assess loneliness. To improve performance, our findings suggest incorporating the diurnal usage data-based app categories features also while developing computational models to assess loneliness.

With the goal to facilitate mental healthcare professionals, we explained the best ML models by SHAP. The analysis showed that the lonely students were more likely to have a lower spending duration per launch of Social Media apps on the weekends. This can be explained by the fact that negative feelings can lead to the launch of Social Media apps to seek social support or to be distracted [29]. We speculate that the lonely students' duration per launch can become lower due to facing negative experiences (e.g., negative content, comparison with others [29]) while using Social Media apps. In the case of the lonely students, we also found that they were likely to use a higher number of Lifestyle apps during the night time period. In that app category, students used apps mostly related to prayer which may also reflect their support-seeking behavior. Our findings through explainable ML which was based on the unobtrusively collected data, are in line with a qualitative study's findings [37] where prayer was found as a coping strategy while going through negative experiences amid the pandemic. Going beyond the study's main focus on loneliness identification, these findings can help mental healthcare professionals to take steps in the interventions.

7 Limitations

Mental health being a taboo topic in Bangladesh [6], the sample size was limited ($N = 105$). This research is expected to generate interest as mental health is a growing research field in developing countries and our findings can facilitate the researchers to develop a better system. Currently, we are conducting a country-wide study and are expecting to come up with a more robust system to more precisely predict loneliness.

8 Conclusion

We present a minimal and real-time system that can identify loneliness by leveraging the instantly (Mean = 0.31 s, SD = 1.1 s) accessed app usage behavioral data. In our study on 105 students of Bangladesh, our developed ML model correctly identified 90.7% lonely students with an F1 score of 82.4%. This shows the efficacy of our minimal system in faster identification of loneliness which can make a worthwhile contribution to minimizing the loneliness rate in low-resource settings.

References

1. World Health Organization (WHO): Mental health atlas 2017. WHO. (2018)
2. Saxena, S., Paraje, G., Sharan, P., Karam, G., Sadana, R.: The 10/90 divide in mental health research: trends over a 10-year period. Br. J. Psychiatry. **188**, 81–82 (2006)
3. Rathod, S., et al.: Mental health service provision in low- and middle-income countries. Health Serv. Insights. (2017)
4. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st NeurIPS. Curran Associates Inc., Red Hook, NY, USA (2017)

5. Mushtaq, R., Shoib, S., Shah, T., Mushtaq, S.: Relationship between loneliness, psychiatric disorders and physical health ? A review on the psychological aspects of loneliness. J. Clin. Diagn. Res. 8, WE01–4 (2014). https://doi.org/10.7860/JCDR/2014/10077.4828
6. WHO: Bangladesh WHO special initiative for mental health situational assessment
7. Kundu, S., et al.: Depressive symptoms associated with loneliness and physical activities among graduate university students in Bangladesh: findings from a cross-sectional pilot study. Heliyon. **7**, e06401 (2021). https://doi.org/10.1016/j.heliyon.2021.e06401
8. Li, Z., Shi, D., Wang, F., Liu, F.: Loneliness recognition based on mobile phone data. In: Proceedings of the 2016 ISAEECE. Atlantis Press, Paris, France (2016)
9. Doryab, A., et al.: Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: Statistical analysis, data mining and machine learning of smartphone and Fitbit data. JMIR MHealth UHealth. **7**, e13209 (2019)
10. Pulekar, G., Agu, E.: Autonomously sensing loneliness and its interactions with personality traits using smartphones. In: 2016 IEEE HI-POCT. IEEE (2016)
11. Wu, C., et al.: Improving prediction of real-time loneliness and companionship type using geosocial features of personal smartphone data. Smart Health (2021)
12. Hays, R.D., DiMatteo, M.R.: A short-form measure of loneliness. J. Pers. Assess. (1987)
13. YourHour - phone addiction tracker & controller. https://play.google.com/store/apps/details?id=com.mindefy.phoneaddiction.mobilepe. Accessed 28 March 2021
14. Ahmed, M.: 86pc university students own smartphones in Bangladesh: Survey. https://en.prothomalo.com/youth/education/86pc-university-students-own-smartphones-in-bangladesh-survey. Accessed 24 Aug 2021
15. Das, R., Hasan, M.R., Daria, S., Islam, M.R.: Impact of COVID-19 pandemic on mental health among general Bangladeshi population: a cross-sectional study. BMJ Open. (2021)
16. Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J.: Machine learning algorithm validation with a limited sample size. PLoS ONE **14**, e0224365 (2019)
17. Ahmed, M.S., Ahmed, N.: Exploring unique app signature of the depressed and non-depressed through their fingerprints on apps. In: Proceeding of the PervasiveHealth'21 (2022)
18. Kursa, M.B., Rudnicki, W.R.: Feature selection with the boruta package. J. Stat. Softw. 36, (2010). https://doi.org/10.18637/jss.v036.i11
19. Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A. (eds.): Feature Extraction: Foundations and Applications. Springer, Berlin (2006)
20. Fusar-Poli, P., McGorry, P.D., Kane, J.M.: Improving outcomes of first-episode psychosis: an overview. World Psychiatry **16**, 251–265 (2017). https://doi.org/10.1002/wps.20446
21. Austin, J., et al.: A smart-home system to unobtrusively and continuously assess loneliness in older adults. IEEE J. Transl. Eng. Health Med. **4**, 2800311 (2016)
22. Coughlan, S.: Loneliness more likely to affect young people. https://www.bbc.com/news/education-43711606 (2018)
23. Guo, Y., Wang, C., Chen, X.: Understanding application-battery interactions on smartphones: a large-scale empirical study. IEEE Access. **5**, 13387–13400 (2017)
24. Rahiem, M.D.H., Krauss, S.E., Ersing, R.: Perceived consequences of extended social isolation on mental well-being: narratives from Indonesian university students during the COVID-19 pandemic. Int. J. Environ. Res. Public Health. **18**, 10489 (2021)
25. Owoyemi, A., Owoyemi, J., Osiyemi, A., Boyd, A.: Artificial intelligence for healthcare in Africa. Front Digit Health. **2**, 6 (2020). https://doi.org/10.3389/fdgth.2020.00006
26. Hunt, M.G., Marx, R., Lipson, C., Young, J.: No more FOMO: limiting social media decreases loneliness and depression. J. Soc. Clin. Psychol. **37**, 751–768 (2018)
27. Zhao, S., et al.: Discovering different kinds of smartphone users through their application usage behaviors. In: Proceedings of the ACM UbiComp'16 (2016)
28. Gao, Y., Li, A., Zhu, T., Liu, X., Liu, X.: How smartphone usage correlates with social anxiety and loneliness. PeerJ **4**, e2197 (2016). https://doi.org/10.7717/peerj.2197

29. Sarsenbayeva, Z., et al.: Does Smartphone Use Drive our Emotions or vice versa? A Causal Analysis. In: Proceedings of the ACM CHI'20 (2020)
30. Mendes, J.P.M., et al.: Sensing apps and public data sets for digital phenotyping of mental health: Systematic review. J. Med. Internet Res. **24**, e28735 (2022)
31. Erzen, E., Çikrikci, Ö.: The effect of loneliness on depression: a meta-analysis. Int. J. Soc. Psychiatry. **64**, 427–435 (2018). https://doi.org/10.1177/0020764018776349
32. Lee, S.L., et al.: The association between loneliness and depressive symptoms among adults aged 50 years and older: a 12-year population-based cohort study. Lancet Psychiatry (2021)
33. UsageStatsManager. https://developer.android.com/reference/android/app/usage/UsageStatsManager. Accessed 15 Sept 2022
34. BANBEIS: bangladesh education statistics 2021 (2022)
35. boruta_py: Python implementations of the Boruta all-relevant feature selection method
36. Zhang, Y., Yang, Y.: Cross-validation for selecting a model selection procedure. J. Econom. **187**, 95–112 (2015). https://doi.org/10.1016/j.jeconom.2015.02.006
37. Finlay, J.M., et al.: Coping during the COVID-19 pandemic: a qualitative study of older adults across the United States. Front. Public Health. **9**, 643807 (2021)
38. Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., Feinstein, A.R.: A simulation study of the number of events per variable in logistic regression analysis. J. Clin. Epidemiol. (1996)
39. Mehrotra, A., et al.: Understanding the role of places and activities on mobile phone interaction and usage patterns. In: Proceedings of the ACM Interaction Mobile Wearable Ubiquitous Technology (2017)
40. Remeseiro, B., Bolon-Canedo, V.: A review of feature selection methods in medical applications. Comput. Biol. Med. **112**, 103375 (2019)
41. Mobile operating system market share Bangladesh. https://gs.statcounter.com/os-market-share/mobile/bangladesh. Accessed 15 Sept 2022

Toward Understanding Users' Interactions with a Mental Health App: An Association Rule Mining Approach

Alaa Alslaity[1(✉)], Gerry Chan[1], Richard Wilson[2], and Rita Orji[1]

[1] Dalhousie University, Halifax, NS, Canada
{Alaa.alslaity,gerry.chan,Rita.Orji}@dal.ca
[2] FeelingMoodie Inc, Moncton, NB, Canada
rw@feelingmoodie.com

Abstract. Mental health apps are gaining increasing research attention. One reason for this is that many users find mental health apps a good alternative for self-management of mental conditions, especially in the last two years when access to physicians was limited because of the COVID-19 pandemic. Despite the existence of several mobile apps targeting mental health, studies show the need to explore and enhance existing mobile health (mHealth) apps to better serve patients and health practitioners. This work aims at analyzing data generated from users of a mobile app to enhance mHealth apps for improving mental health. Particularly, this paper aims to extract knowledge about the relationship between different activities (e.g., sport, home, school, etc.) that affect users' moods. To achieve this goal, an association rule mining technique was applied on a dataset collected in the wild from 232 users of a mental health app called the FeelingMoodie app. They used the app from September 2021 to May 2022. Our results revealed interesting associations between various daily life activities. Based on these association rules, we provide insights and recommendations for building better mHealth apps and a more personalized user experience.

Keywords: Mental health · Mobile health app · mHealth · Mood tracker · Mobile app · Association Rule Mining

1 Introduction

Recent years have witnessed significant adoption of technology in the health domain [4, 26]. Advanced technologies are nowadays an essential tool in the toolset of health service providers and patients. These technologies allowed for the delivery of health care services on a larger scale and with more efficiency. Among these technologies are mobile health (mHealth) apps, which are defined as "medical or public health practice supported by mobile devices" [20]. mHealth apps have shown strong evidence in supporting conventional health systems by augmenting diagnosing and treatment processes and increasing individuals' participation in managing their mental health. They have been shown to be useful for a wide range of health areas. Due to its importance, mental health apps comprise about one-third of disease-specific mHealth apps [29].

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 477–495, 2023.
https://doi.org/10.1007/978-3-031-34586-9_32

Despite their prevalence, existing studies have shown that to be effective, mHealth apps need to be explored and enhanced so that they provide the best assistance for patients and health practitioners. For instance, existing studies show that mHealth apps need to be personalized, interactive, and easy to use. The available mHealth apps collect a significant amount of data, which can be explored to generate insights that can be used to advance these apps. Various data analytic approaches for processing data and generating insights from data have been explored by knowledge discovery and data mining (DM) researchers. Knowledge Discovery (KD) is defined as the process of analyzing and discovering interesting knowledge and patterns from a large amount of data [19]. Data Mining (DM) is the core process of knowledge discovery from databases [6]. It is used to extract meaningful information and infer relationships among variables, through classification, clustering, and association rule mining [5]. Association Rule Mining (ARM), which is of special importance for this study, is one of the most common techniques of data mining. It was introduced in the early 1990s [1], and it aims to infer interesting correlations and associations, and frequent patterns among sets of items in the data. ARM is widely used in different domains, including market, risk management, medical diagnosis, bio-medical literature, protein sequences, and inventory control [35].

This work aims to use existing data to enhance mHealth apps for improving mental health. Particularly, the goal of this research is to extract knowledge about the relationship between different activities that affect users' moods. To achieve this goal, ARM techniques were applied on a dataset generated by users of a mental health app, called the *Feeling Moodie* (*Moodie* for short) app [3, 22]. Users used the *Moodie* app to log their moods along with associated activities that users believe affect their mood. The app allows users to select moods that belong to five main categories (Good, Mad, Sad, Rad, and Neutral). Then, users can select one or more activities (e.g., family, event, friends, sports) that might have affected their moods. The data contains 2336 records collected from 232 users from September 2021 to May 2022, and it contains several features, including moods and activities, which are related to our study.

This work is part of ongoing research toward developing an adaptive and evidence-based mental health app. Identifying the relationship between moods and environmental aspects can help understand what factors (both negative and positive) and patterns impact people's moods and wellbeing. As a first step in designing an AI-based adaptive mental health app, this work aims to explore the relationship between factors affecting moods and how this relationship varies based on users' moods to inform the design of better mental health apps. To achieve our goal, we adopt the association rule mining concepts and applied them to uncover the relationships between activities. ARM explores the relationship between unrelated data. For instance, a supermarket can use data about customers' purchases and ARM to identify which products are frequently bought together. Association rules are if/then statements that identify the relationship between data elements [24], and they are used to identify the objects that frequently happen (or used) together. In this work, we used the same technique to uncover the relationship between different activities (or factors) that may affect users' moods. That is, we explore the relationship between the activities that lead the user to be in a particular mood.

The quantitative analysis of the data revealed that Home-, Work-, Relaxation-, and Family-related activities are the most common factors that affect users' moods, and Chill emerged as the most common mood. The current work adds to the ongoing efforts to understand how technology can be used to enhance positive wellbeing and how humans interact with mood-tracking apps for improving mental wellness. The analysis also revealed several association rules that describe the relation and association between different activities. This research contributes to the ongoing efforts to understand how self- and mood-tracking apps can be used to change negative to positive moods, as well as restore positive moods while reducing negative ones. This research also provides insights and recommendations for designing adaptive user-centric mental health apps.

2 Background and Related Work

This section introduces association rule mining and its applications. Then, it discusses related work on mHealth apps.

2.1 Association Rule Mining

Association Rule Mining (ARM) is an effective data mining technique that uses rule-based machine learning methods to discover relations between items in a dataset [2, 5]. Given transactions with a variety of items, ARM is meant to explore the rules that determine how certain items are connected. Association rules were firstly defined in the early 1990s by Agrawal et al. [1], as follows: let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items, where items could be literals, shopping items, images, etc. Any subset $X \in I$ is called itemset. Let R be a table with a set of transactions t involving elements that are a subset of I. An association rule is an expression of the form $X \rightarrow Y$, where X and Y are itemsets and $X \cap Y = \emptyset$. X is called the antecedent (or body) of the rule, and Y is the consequent (or head) of the rule. An ARM algorithm usually mines several rules. The number of the mined association rules depends on the size of the dataset, and two essential measures: *Support* (S), and *Confidence* (C) [1], which are defined as follows [34]. *Support* is the percentage of transactions (or records) that contain X and Y to the total number of transactions $|X|$. *Confidence* is the percentage of the number of transactions that contain X and Y to the total number of records that contain X. Confidence shows the strength of the association rules. For instance, if the confidence of the association rule $X \rightarrow Y$ is 0.8, it means that 80% of the transactions that contain X also contain Y [35].

The process of mining association rules involves finding rules that satisfy minimum support and minimum confidence. These thresholds are domain-dependent, they are specified by the user, and they are used to drop less interesting and less important rules. Sometimes even the association rules generated from not so frequent itemsets (low support value) are still important. For instance, some items are not purchased so often because they are very expensive, consequently, they are not purchased as often as the threshold required. However, the association rules between those expensive items are as important as other frequently bought items to the retailer. Once the list of the most important rules is identified, it can be ordered based on *Lift* value (a measure of the importance of a rule), which helps select rules with high predictive power [34]. For example, the lift of α for the association rule $X \rightarrow Y$ tells us that Y is α times more likely to be bought by the customers who also buy X compared to the default likelihood sale of Y. Many algorithms for generating association rules were presented over time, such as Eclat, FP-Growth, and Apriori algorithms [6]. The most commonly used approach for finding association rules is based on the Apriori algorithm [35].

ARM is one of the most important data mining tasks. It has been extensively studied and applied in marketing, where it is known as the "market basket analysis". For instance, in a bookstore, association rules discovered from the transaction database can be used to rearrange the presentation of books on the shelves such that books that are found to be bought together are placed close to each other. Also, association rules are used for building marketing strategies. For instance, if the ARM process revealed that item Y is bought with item X 60% of the time, it indicates that promoting X can increase the sales of Y. In addition to this typical use of ARM, it has been also introduced to a wide array of applications, such as: classifying the student based on their performance in academics [6], to find the best combination of courses based on users' enrolment data [9], classifying text documents in associating terms of text categories [27, 39], and in transportation domain to explore the causes of accidents and maintenance issues [2, 28].

In addition, ARM techniques can be applied in the health domain (which is the focus of our research) to achieve various goals. For instance, Ribeiro et al., [34] proposed a method based on association rule-mining to enhance the diagnosis of medical images. Pan et al., [32] presented a solution using association rules to relate objects and categories of brain tumors. Wang et al., [38] investigated using ARM techniques to discover patterns of interest in a dataset containing digital mammographic images and textual reports of radiologists. Other researchers [30] deployed ARM and heart disease data to predict healthy and sick heart. Several researchers have also explored the use of ARM for disease and health issues [10, 25], such as cardiac diseases [31], diabetes [15, 16], cancer analysis [33], and explore epidemic and pandemics (e.g., dengue fever [11, 37] and virus outbreaks such as Ebola virus [18], and COVID-19 [23, 36]).

2.2 Mental Health Apps

The use of mental health apps has been shown to improve health behaviours and help with the regulation of moods. A recent review of the literature on mHealth apps that foster emotion regulation, positive mental health, and wellbeing found that although there is emerging evidence showing the benefits of these apps for improving mental health and wellbeing, very few apps are targeted at promoting emotion regulation. The researchers concluded that future mHealth apps may consider the inclusion of features that promote emotion regulation. One of the aims of the present research study is to make it easier for users to identify connections in their moods and help with emotion regulation, and coping with the ups and downs of life.

There has also been research on how people's moods are correlated with activities of daily living [the fundamental skills required to independently care for oneself such as eating, shopping, housecleaning, and communication with others. For example, Chan et al. [13] investigated the relationship between daily mood changes and one's personal characteristics, demographic factors, and daily health behaviours. For 30 days, 130 users completed a program called "ClickDairy" which allows users to enter their daily health-related behaviours. Results showed that a user's mood can be associated with health behaviours and daily life activities. The same conclusion has also been demonstrated by other researchers [3, 12]. Results also showed that users experience better moods on the weekends and users who perform more exercise experience better moods compared to users who do not perform an exercise. The quality of sleep was also related to mood fluctuations from day to day – better-quality sleep leads to the experience of a more positive mood the following day. In a different study, Bakker and Rickard [17] evaluated a self-reflection focussed app called "MoodPrism" for improving one's mental health and wellbeing. Results showed that users who had more positive and engaging experiences using the app experienced greater decreases in depression and anxiety. Results also showed that users who already had knowledge of mental health issues were more likely to use MoodPrism over the longer term.

More recently, Huberty et al. [21] examined the effectiveness of the "Calm" app for reducing stress and improving mindfulness and self-compassion among college students. Results showed that the use of the Calm app helped reduce stress, and self-compassion, and improve sleep quality. However, results did not show any association between the use of the app and changes in health-related behaviours such as alcohol consumption, physical activity, or healthy eating. The researchers concluded that there is a lack of research evaluating the impact of mindfulness mediation mobile apps on health behaviour outcomes and the short- and long-term effects.

Similarly, Cho et al. [14] conducted a 1-year pilot study to evaluate the effectiveness of a smartphone app called "Circadian Rhythm for Mood" for helping patients with mood disorders prevent the reoccurrence of mood episodes. Results showed that the total number of mood episodes was fewer and shorter for patients who used the app. Positive changes in health behaviour were observed and the app was found to be effective in preventing the reoccurrence of mood disorders, improving prognosis, and promoting better health behaviours.

Finally, Athanas et al. [8] investigated the effects of immediate and long-term use of a guided meditation and mindfulness app called "Stop, Breathe & Think (SBT)" on users' emotional states. To explore the long-term effects, the changes in the user's basal emotional state were assessed before they completed the guided meditation activity. Results showed repeated engagements with the SBT app are associated with an improvement in users' emotional states over time. Results also showed that after using the app for an extended period, users who felt sad at the beginning of the intervention became happier. The researchers concluded that repeated use of the SBT app can change a user's emotional state from negative to positive and suggest that more elaborate studies are needed to better understand the potential benefits of these apps to reduce the cost of healthcare services.

While there are many studies on the use of mHealth apps for reducing stress and improving health and wellbeing, and how moods are associated with people's daily life activities, there is a lack of work on using association rule mining to uncover how different activities are associated, the result of which could be used to inform the development of mHealth apps that are more adaptive, and easy to use. In this paper, we describe how we adopted the ARM concepts to identify relationships between activities that affect users' moods.

3 Method

This section discusses the research method we used to achieve our goals. First, it describes the *Moodie* app and its features. Then, it discusses the implementation of the Apriori algorithm (the association rule mining algorithm we used in our study).

3.1 The Moodie Mental Health App

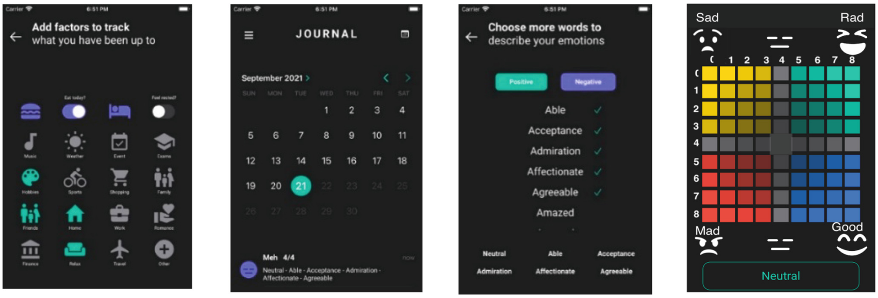
The *Feeling Moodie* (or *Moodie*) app [22] is a mental health app that tracks mood and health-related information. It is a HIPPA (Health Insurance Portability and Accountability) [7] and PIPEDA (Personal Information Protection and Electronic Documents Act) [40] compliant. The app allows users to track their moods by reporting their moods and associated activities (Family, Events, Friends, Sports, Travel, Romance, Finance, Shopping, Exam, Relax, Hobbies, Music, Work, Sleep, Home, Food, Weather, and Other). These activities are the most common activities, and they were selected based on an exploration of related apps. Users can identify their moods using a mood grid designed around four moods dimensions: Great (Happiness or good mood), Chill (being in a calm state, or to relax completely or being relaxed), Meh (indifference, boredom, or a lack of enthusiasm), and Sad (feelings of unhappiness and low mood). These four moods were selected as general moods representing a wide range of sub-moods. To make sure that these moods are clear, the app present the wide range of submoods as a grid, which makes moods clearer to users. The sub moods explain the general moods and provide more fine-rained categories. The Moodie app also allows users to enter free-text journaling entries to talk about their moods and associated activities.

Moodie app has four main features: (1) *Mood Tracking* where users can enter their moods and associated activities, (2) *Mind* Fitness Plan, which provides assistance for users to change their mood to a better state using cognitive behavioral therapy [11] treatment, (3) Journaling, which allows users to track events and situations that might affect their mood and (4) Visualization where users can see historical data and track their moods, which help users' and health care practitioners to recognize emerging trends and respond rapidly. Figure 1 depicts screenshots of the *Moodie* app.

3.2 Implementation

To analyze the association between different activities that affect users' moods, we applied the *Apriori* algorithm, a well-known and one of the most widely used ARM algorithms. It is mainly used for finding frequent items over transactional data sources [27]. Apriori algorithm takes a set of itemsets and tries to find the most common subsets. To do so, Apropri algorithm uses a "bottom up" approach where frequent subsets are extended one item at a time in a step known as candidate generation, which is based on the minimum support value. Apriori uses breadth-first search and a tree structure to count candidate itemsets efficiently. We used Apriori algorithm because it is easy to implement, and works efficiently in relatively small dataset [5]. So, it is suitable for our study.

All our experiments were executed in a laptop with Intel Core i7 CPU, at 2.1 GHz (4CPUs), and Windows 10 as the operating system. Python 3.7 (64-bit) was used to



(a) Activity tracking. (b) Journaling. (c) Emotion tracking. (d) Mood grid.

Fig. 1. Main features of the Moodie App[1]

implement the ARM algorithm. We set the minimum support (0.2) and the minimum confidence (0.8), to obtain all frequent itemsets and filter out the strong association rules based on confidence. Then, the lift value was considered to sort the rules based on their importance.

3.3 Dataset

The data set used in this study was obtained from users who used the Version II of the *Moodie* app. The data contains several features, including moods and activities, which are related to our study. Due to privacy concerns, users demographics were not collected and, therefore, it is not considered in our study. As mentioned above, users can use the *Moodie* app to log their moods along with associated activities that users believe that they affect their entered mood. The app allows users to select moods that belong to five main categories (Good, Mad, Sad, Rad, and Neutral). Then, users can select one or more activities that might have affected their moods. Particularly, eighteen different activities were provided to users: family, events, friends, sports, travel, romance, finance, shopping, exam, relax, hobbies, music, work, sleep, home, food, weather, and others. The data contains 2,336 records collected from 232 users from September 2021 to May 2022. Users logged their moods and activities as they wished and at the convenience of their schedule. According to previous studies [5], most of the research in ARM for health informatics uses small to medium datasets, containing less than 2,000 records. So, the number of records available in our dataset is adequate to mine association rules.

4 Results

This section discusses the results based on the data collected using the *Moodie* app. First, it shows a descriptive analysis of mood and activities distribution based on the overall data (including all the moods and activities). Then, it represents the association rule analysis of the overall data, followed by association rules per mood.

---

[1] AppAdvice: https://appadvice.com/app/feeling-moodie/1581336127.

4.1 Descriptive Analysis

A cross-tabulation of moods and activities is shown in Table 1, where the columns show moods, and the rows represent activities. The bolded-text between brackets values indicate the least frequent activity for each mood, while the bolded-text without brackets are the most frequent activities. As shown in Table 1, Finance is the least frequent activity for Rad, Neutral, and Sad; Travel is the least frequent activity for Good and Neutral; and Shopping is the least frequent for Mad. On the other hand, Food is the most frequent activity for all moods according to users' self-reports. This means that Food is the most frequent activity overall, as it is depicted in Fig. 2, which shows the overall frequency for each activity in the whole dataset. As Fig. 2 shows, Food is the most frequently selected activity, followed by Sleep, Home, and Relax. In contrast, Travel is the least frequently selected activity, while Finance and Shopping is the second and third least frequently selected activities, consequently.

Table 1. Frequency of moods and activities

Activity	Rad	Good	Neutral	Mad	Sad
Family	132	214	78	19	73
Event	40	67	23	5	26
Friends	107	168	68	19	61
Sprots	77	113	43	7	38
Travel	18	**(17)**	**(16)**	5	12
Romance	73	136	45	7	46
Finance	**(13)**	25	**(16)**	7	**(12)**
Shopping	43	31	17	**(2)**	17
Exams	71	133	82	16	100
Relax	133	272	93	10	71
Hobbies	59	71	27	4	22
Music	75	107	43	13	56
Work	100	171	91	43	101
Sleep	293	427	164	44	118
Other	45	54	31	16	63
Home	177	319	146	33	122
Food	**348**	**604**	**267**	**85**	**245**
Weather	88	117	58	18	48

4.2 Association Rule Mining

This section shows the association rules mined by the Apriori algorithm, setting the minimum support to (0.2) and minimum confidence (0.8) in order to filter the most significant rules. After getting the most significant rules, we sorted them based on the lift value, which measures the importance and interestingness of the rule [6]. For clarity purposes, the results presented in this section show the ten most important rules. A complete list of the rules can be provided as a supportive material.

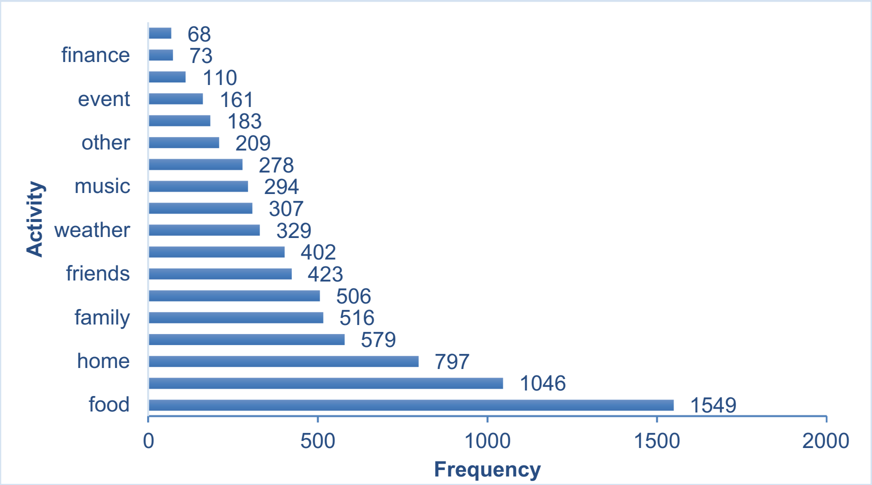


Fig. 2. Frequency of activities over all data

Overall Data Analysis

This section presents the association rules for the whole dataset. The total number of rules generated was 63 with confidence $>80\%$. Table 2 shows the top ten association rules. As the table shows, the top rule indicates that moods that are affected by Romance- and Sport-related activities, are more likely affected by Family-related activities. It also shows that Sleep is associated with six itemsets. This is not surprising given that Sleep is the second most common activity (as shown in Fig. 2).

Mood-Based Analysis

This section presents the association rules for each mood. Tables 3, 4, 5, 6 and 7 depict the association rules per mood. That is, each table represents the association rules mined using only the records of the corresponding mood. Table 3 shows the association rules for Neutral mood. Our analysis revealed a total of 100 rules related to Neutral mood. The top rules show that Romance-related activities are highly associated with Weather, and Hobbies activities. This rule has (8.34) lift value with high confidence, which means that whenever Weather- and Hobbies-related activities made users be in a neutral mood, then there is a very high chance that a Romance-related activity is also affecting their

Table 2. Association rules based on the whole dataset

#	Rule	Confidence	Lift
1	{‘Romance’, ‘Sports’} → {‘Family’}	0.808	3.655
2	{‘Shopping’, ‘Relax’} → {‘Sleep’}	0.923	2.061
3	{‘Shopping’, ‘Food’} → {‘Sleep’}	0.882	1.968
4	{‘Hobbies’, ‘Sports’} → {‘Sleep’}	0.873	1.948
5	{‘Shopping’, ‘Home’} → {‘Sleep’}	0.841	1.876
6	{‘Music’, ‘Hobbies’} → {‘Sleep’}	0.806	1.799
7	{‘Music’, ‘Romance’} → {‘Sleep’}	0.803	1.793
8	{‘Shopping’, ‘Sleep’} → {‘Food’}	0.965	1.454
9	{‘Friends’, ‘Sleep’} → {‘Food’}	0.948	1.429
10	{‘Hobbies’, ‘Sleep’} → {‘Food’}	0.944	1.423

Table 3. Association rules for “Neutral” mood

#	Association Rule	Confidence	Lift
1	{‘Weather’, ‘Hobbies’} → {‘Romance’}	0.900	8.340
2	{‘Hobbies’, ‘Sports’} → {‘Friends’}	0.900	5.519
3	{‘Shopping’, ‘Friends’} → {‘Family’}	1.000	5.346
4	{‘Weather’, ‘Romance’} → {‘Friends’}	0.867	5.315
5	{‘Romance’, ‘Work’} → {‘Friends’}	0.846	5.189
6	{‘Shopping’, ‘Family’} → {‘Friends’}	0.833	5.110
7	{‘Romance’, ‘Sports’} → {‘Friends’}	0.813	4.983
8	{‘Romance’, ‘Sleep’} → {‘Friends’}	0.808	4.953
9	{‘Hobbies’, ‘Sleep’} → {‘Friends’}	0.800	4.906
10	{‘Friends’, ‘Weather’} → {‘Family’}	0.905	4.837

moods. Although As Fig. 2 shows, Food is the most frequently selected activity, followed by Sleep, Home, and Relax. In contrast, Travel is the least frequently selected activity, while Finance and Shopping is the second and third least frequently selected activities, consequently.

4.3 Association Rule Mining

This section shows the association rules mined by the Apriori algorithm, setting the minimum support to (0.2) and minimum confidence (0.8) in order to filter the most significant rules. After getting the most significant rules, we sorted them based on the lift value, which measures the importance and interestingness of the rule [6]. For clarity

Table 4. Association rules for "Good" mood

#	Association Rule	Confidence	Lift
1	{'Exams', 'Hobbies'} → {'Food'}	1.000	1.377
2	{'Music', 'Hobbies'} → {'Food'}	1.000	1.377
3	{'Other', 'Sleep'} → {'Food'}	1.000	1.377
4	{'Shopping', 'Sleep'} → {'Food'}	1.000	1.377
5	{'Hobbies', 'Sleep'} → {'Food'}	0.980	1.349
6	{'Romance', 'Hobbies'} → {'Food'}	0.957	1.318
7	{'Exams', 'Sleep'} → {'Food'}	0.953	1.313
8	{'Friends', 'Sleep'} → {'Food'}	0.951	1.310
9	{'Other', 'Family'} → {'Food'}	0.947	1.305
10	{'Shopping', 'Home'} → {'Food'}	0.947	1.305

Table 5. Association rules for "Sad" mood

#	Association Rule	Confidence	Lift
1	{'Event', 'Sports'} → {'Friends'}	0.900	6.374
2	{'Exams', 'Sports'} → {'Friends'}	0.867	6.138
3	{'Romance', 'Sports'} → {'Friends'}	0.846	5.992
4	{'Romance', 'Sports'} → {'Family'}	1.000	5.918
5	{'Family', 'Sports'} → {'Friends'}	0.800	5.666
6	{'Event', 'Sports'} → {'Family'}	0.900	5.326
7	{'Friends', 'Sports'} → {'Family'}	0.842	4.983
8	{'Event', 'Romance'} → {'Exams'}	0.917	3.960
9	{'Event', 'Friends'} → {'Exams'}	0.833	3.600
10	{'Music', 'Work'} → {'Sleep'}	0.833	3.051

purposes, the results presented in this section show the ten most important rules. A complete list of the rules can be provided as a supportive material.

Although Table 1 shows that Friends-related activities are not among the most common activities related to Neutral mood, the ARM results show that Friends-related activities are associated with seven itemsets, with high lift values, as shown in Table 3 (Rules 2, and 4 to 9). This means that whenever any of these seven itemsets made the user be in a Neutral mood, then it is almost five times more probable that their moods have also been affected by Friends-related activities.

With regards to the Good mood, Table 4 shows the top ten rules among a total of 54 association rules. Surprisingly, all the top ten association rules show the relation

Table 6. Association rules for "Mad" mood

#	Association Rule	Confidence	Lift
1	{'Friends', 'Hobbies'} → {'Relax'}	1.000	14.300
2	{'Exams', 'Family'} → {'Music'}	1.000	11.000
3	{'Romance', 'Sleep'} → {'Music'}	1.000	11.000
4	{'Weather', 'Sleep'} → {'Music'}	0.857	9.429
5	{'Exams', 'Music'} → {'Family'}	1.000	7.526
6	{'Finance', 'Home'} → {'Family'}	1.000	7.526
7	{'Friends', 'Music'} → {'Family'}	1.000	7.526
8	{'Family', 'Romance'} → {'Friends'}	1.000	7.526
9	{'Friends', 'Romance'} → {'Family'}	1.000	7.526
10	{'Music', 'Work'} → {'Family'}	1.000	7.526

between an itemset and Food activity. Also, we notice that the confidence values of these association rules are very high (>94%), with 100% confidence of the top four rules.

A total of 51 association rules were found in the Sad-related data. Table 5 shows the top association rules related to Sad mood. The top three rules show that if the users' mood is Sad because of {'Event', 'Sports'}-, {'Exams', 'Sports'}, or {'Romance', 'Sports'}-related activities, then it is highly likely that Friends-related activities have also affected users' moods and made them feel sad.

Table 6 summarizes the top association rules among 63 rules related to the Mad mood. It can be noticed that these rules have a very high confidence (100% for most of them). Also, the lift values for these rules are high. For instance, Relax activity is 14.3 times more likely to cause a Mad mood if Friends and Hobbies related activities also made users feel Mad.

Finally, the ARM revealed a total of 141 rules related to Rad mood. The top ten rules are shown in Table 7. The table shows that Music and Sports are 5.76 and 5.61 times, consequently, more likely to make users feel Rad if Exams- and Hobbies-related activities made them Rad. It also shows that Family is the consequent of five antecedent itemsets (rules 4–8). These five associations indicate that Family is an important activity that leads to Rad mood. Although it is not the most common activity in rad mood (as shown in Table 1).

5 Discussion and Future Work

In the previous section, we presented the results obtained from our ARM analysis. These rules can be used to develop mHealth apps that are more adaptive, and easy to use. Particularly, the benefits of these association rules can be summarized as follows:

Better Understanding of Users. The mined association rules allow designers of mHealth apps and researchers to better understand factors that are associated with users' moods by

Table 7. Association rules for "Rad" mood

#	Association Rule	Confidence	Lift
1	{'Exams', 'Hobbies'} → {'Music'}	0.846	5.765
2	{'Exams', 'Hobbies'} → {'Sports'}	0.846	5.615
3	{'Shopping', 'Music'} → {'Weather'}	0.923	5.360
4	{'Shopping', 'Friends'} → {'Family'}	1.000	3.871
5	{'Shopping', 'Music'} → {'Family'}	0.846	3.276
6	{'Romance', 'Sports'} → {'Family'}	0.824	3.188
7	{'Shopping', 'Weather'} → {'Family'}	0.813	3.145
8	{'Exams', 'Weather'} → {'Family'}	0.810	3.134
9	{'Shopping', 'Sports'} → {'Home'}	0.917	2.646
10	{'Shopping', 'Relax'} → {'Home'}	0.909	2.625

discovering more activities that might affect the user's mood but are not mentioned by the user. For example, based on the association rules ({'Romance', 'Sports'} → {'Family'}) mentioned in Table 2, if the user mentioned that Romance- and Sport-related activities only, the system understands that the user's mood could also be affected by Family-related activities. Therefore, the app may recommend interventions related to family activities as well.

Enhancing Credibility and Usability. The rules can be used to reduce the steps and inputs required from users. For example, instead of asking the user to continuously enter as many activities as possible, the system may ask for 2–4 inputs. Based on these inputs, the system can predict other activities of various kinds based on the provided rules. For instance, suppose that the user selected {Romance, Sport, and Hobbies} as the activities that caused their mood. If we look at Table 2, searching for the rules that have at least two of these three activities, we will find 3 rules. Accordingly, the system can infer that in addition to these three activities, Sleep, and Family are most likely affecting the user's mood as well. Adding these capabilities to the apps will increase their credibility and usability.

Personalization and Adaptation. The app can also use our results to personalize the list of activities and make it more adaptive. For instance, the order of the activities in the list can be changed based on their importance (how they are related) to the activities selected by users, or the activities can be presented as groups. This adaptability will in turn enhance the usability and credibility of the app.

As shown in Sect. 4.2, the results revealed interesting patterns that can inform the development of an adaptive AI-based mental health app. APPENDIX 1 summarizes the results presented in the previous section and acts as a reference for designers and researchers. Below is a summary of the key observations and recommendations/guidelines for designing mHealth apps:

- Food, Sleep, and Home are the most common activities, while Travel, Finance, and Shopping, events are the least common activities. It seems that the COVID-19 pandemic has had an impact on users' selections since, during the pandemic, people spent more time at home because of isolation and work from home. Also, during that time shopping activities were limited, and many people switched to delivery service options. To best understand the impact of the context on users' selection of activities, our data collection is continuing.
- Food is the most common activity in the overall dataset. Considering this observation in isolation reveals that a mental health app should emphasize food-related activities. However, the ARM results (Sect. 4.2) revealed that Food activities are associated with other activities. For instance, the association rules of the overall data (Table 2) shows that {'Shopping', 'Sleep'}, {'Friends', 'Sleep'}, and {'Hobbies', 'Sleep'} are all associated with Food activities. Also, other activities are associated with Food. For instance, Table 2 shows that if the user selected 'Shopping' and 'Food' activities, then it is more likely the user will select 'Sleep' activity to be associated with their moods. Therefore, designers should not focus only on the absolute popularity of the activities, but also consider the associated activities.
- Overall, Travel is the least frequent activity. Also, the top association rules did not reveal any association between Travel and other activities. However, it is important to mention that the data collection was done during the COVID-19 pandemic when travel was limited in most countries. So, users' activities during this time might be affected, and therefore their selection of activities. It will be important to do further studies as the pandemic situation shifts to an endemic.
- Friend-related activities are associated with neutral mood, although it is not among the most common activities related to neutral mood. Our results (see Table 3) indicate that Friends-related activities are associated with seven of the top itemsets (with a lift value of more than 5). This means that whenever the users indicate that any of these seven itemsets made the user feels Neutral, then it is almost five times more probable that their moods have also been affected by Friends-related activities.

5.1 Limitations and Future Work

Despite the interesting results and insights revealed by this study, this study is based on data collected in the wild, demographical information was not collected due to privacy reasons. Also, it is worth mentioning that the data were collected during the COVID-19 pandemic, we believe that it might have affected users' daily life activities (e.g., travel). Therefore, as a future work, we will conduct a subsequent study to expand on this study, investigate, and compare the association rules after the pandemic. Again, although, Mad-related association rules has a very high confidence (100%). However, Mad is the least frequent mood in the dataset (only 353 records). So, we need to collect more data related to this mood and repeat the analysis.

6 Conclusion

This paper presented our ongoing effort toward providing adaptive and personalized mental health apps. In this research, we used a dataset containing 2,336 records collected from 232 users of a mental health app in the wild, from September 2021 to May 2022. The mental health app is called the *FeelingMoodie* app. We used this data to extract knowledge about the relationship between different daily life activities that affect individuals' moods. To achieve our goals, we applied the association rule mining techniques on the dataset. Our results revealed interesting associations between various daily life activities, and these associations vary based on mood. The paper also shows how to use these rules to enhance mental health apps and provides insights and recommendations for building better mHealth apps and a more personalized user experience.

APPENDIX 1. Summary of the Association Rules

Rule		Mood
Antecedent	Consequent	
{'Romance', 'Sports'}	{'Family'}	Overall
{'Event', 'Friends'}	{'Exams'}	Sad
{'Event', 'Romance'}	{'Exams'}	Sad
{'Event', 'Sports'}	{'Family'}	Sad
	{'Friends'}	Sad
{'Exams', 'Family'}	{'Music'}	Mad
{'Exams', 'Hobbies'}	{'Food'}	Good
	{'Music'}	Rad
	{'Sports'}	Rad
{'Exams', 'Music'}	{'Family'}	Mad
{'Exams', 'Sleep'}	{'Food'}	Good
{'Exams', 'Sports'}	{'Friends'}	Sad
{'Exams', 'Weather'}	{'Family'}	Rad
{'Family', 'Romance'}	{'Friends'}	Mad
{'Family', 'Sports'}	{'Friends'}	Sad
{'Finance', 'Home'}	{'Family'}	Mad
{'Friends', 'Hobbies'}	{'Relax'}	Mad
{'Friends', 'Music'}	{'Family'}	Mad

(*continued*)

(*continued*)

Rule		Mood
Antecedent	Consequent	
{‘Friends’, ‘Romance’}	{‘Family’}	Mad
{‘Friends’, ‘Sleep’}	{‘Food’}	Good
	{‘Food’}	Overall
{‘Friends’, ‘Sports’}	{‘Family’}	Sad
{‘Friends’, ‘Weather’}	{‘Family’}	Neutral
{‘Hobbies’, ‘Sleep’}	{‘Food’}	Good
	{‘Friends’}	Neutral
	{‘Food’}	Overall
{‘Hobbies’, ‘Sports’}	{‘Sleep’}	Overall
	{‘Friends’}	Neutral
{‘Music’, ‘Hobbies’}	{‘Food’}	Good
{‘Music’, ‘Hobbies’}	{‘Sleep’}	Overall
{‘Music’, ‘Romance’}	{‘Sleep’}	Overall
{‘Music’, ‘Work’}	{‘Family’}	Mad
	{‘Sleep’}	Sad
{‘Other’, ‘Family’}	{‘Food’}	Good
{‘Other’, ‘Sleep’}	{‘Food’}	Good
{‘Romance’, ‘Hobbies’}	{‘Food’}	Good
{‘Romance’, ‘Sleep’}	{‘Friends’}	Neutral
	{‘Music’}	Mad
{‘Romance’, ‘Sports’}	{‘Family’}	Sad
	{‘Family’}	Rad
	{‘Friends’}	Neutral
	{‘Friends’}	Sad
{‘Romance’, ‘Work’}	{‘Friends’}	Neutral
{‘Shopping’, ‘Family’}	{‘Friends’}	Neutral
{‘Shopping’, ‘Food’}	{‘Sleep’}	Overall
{‘Shopping’, ‘Friends’}	{‘Family’}	Neutral
	{‘Family’}	Rad
{‘Shopping’, ‘Home’}	{‘Sleep’}	Overall
	{‘Food’}	Good
{‘Shopping’, ‘Music’}	{‘Family’}	Rad
	{‘Weather’}	Rad
{‘Shopping’, ‘Relax’}	{‘Home’}	Rad

(*continued*)

(*continued*)

Rule		Mood
Antecedent	Consequent	
	{'Sleep'}	Overall
{'Shopping', 'Sleep'}	{'Food'}	Good
	{'Food'}	Overall
{'Shopping', 'Sports'}	{'Home'}	Rad
{'Shopping', 'Weather'}	{'Family'}	Rad
{'Weather', 'Hobbies'}	{'Romance'}	Neutral
{'Weather', 'Romance'}	{'Friends'}	Neutral
{'Weather', 'Sleep'}	{'Music'}	Mad

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. ACM SIGMOD Rec. **22**(2), 207–216 (1993). https://doi.org/10.1145/170036.170072
2. Ait-Mlouk, A., Gharnati, F., Agouti, T.: An improved approach for association rule mining using a multi-criteria decision support system: a case study in road safety. Eur. Transp. Res. Rev. **9**(3), 1–13 (2017). https://doi.org/10.1007/S12544-017-0257-5/TABLES/9
3. Alslaity, A., Chan, G., Wilson, R., Orji, R.: Insights from longitudinal evaluation of moodie mental health app. In: Conference on Human Factors in Computing Systems – Proceedings (2022). https://doi.org/10.1145/3491101.3519851
4. AlSlaity, A., Suruliraj, B., Oyebode, O., Fowles, J., Steeves, D., Orji, R.: Mobile applications for health and wellness: a systematic review. In: Proceedings of the ACM on Human-Computer Interaction, vol. 6, no. EICS, pp. 1–29 (2022). https://doi.org/10.1145/3534525
5. Altaf, W., Shahbaz, M., Guergachi, A.: Applications of association rule mining in health informatics: a survey. Artif. Intell. Rev. **47**(3), 313–340 (2017). https://doi.org/10.1007/S10462-016-9483-9/FIGURES/6
6. Angeline, D.M.D.: Association rule generation for student performance analysis using Apriori algorithm. SIJ Trans. Comput. Sci. Eng. Appl. (CSEA) **1**(1), 12–16 (2013)
7. Atchinson, B.K., Fox, D.M.: The politics of the health insurance portability and accountability act. Health Affairs (Proj. Hope) **16**(3), 146–150 (1997). https://doi.org/10.1377/HLTHAFF.16.3.146
8. Athanas, A.J., et al.: Association between improvement in baseline mood and long-term use of a mindfulness and meditation app: observational study. JMIR Ment. Health **6**, 5 (2019). https://doi.org/10.2196/12617
9. Aher, S.B., Lobo, L.M.R.J.: Combination of clustering, classification & association rule based approach for course recommender system in e-learning. Int. J. Comput. Appl. **39**(7), 8–15 (2012). https://doi.org/10.5120/4830-7087
10. Berka, P., Rauch, J.: Mining and post-processing of association rules in the atherosclerosis risk domain. In: Khuri, S., Lhotská, L., Pisanti, N. (eds.) ITBAM 2010. LNCS, vol. 6266, pp. 110–117. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15020-3_11

11. Buczak, A.L., Koshute, P.T., Babin, S.M., Feighner, B.H., Lewis, S.H.: A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. BMC Med. Inform. Decis. Mak. **12**(1), 1–20 (2012). https://doi.org/10.1186/1472-6947-12-124/FIGURES/24
12. Chan, G., Alslaity, A., Wilson, R., Orji, R.: Exploring variance in users' moods across times, seasons, and activities: a longitudinal analysis. In: MobileHCI (2022)
13. Chan, T.C., Yen, T.J., Yang Chih, F., Hwang, J.S.: ClickDiary: online tracking of health behaviors and mood. J. Med. Internet Res. **17**(6), e147 (2015). https://doi.org/10.2196/JMIR.4315
14. Cho, C.H., Lee, T., Kim, M.G., In, H.P., Kim, L., Lee, H.J.: Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: prospective observational cohort study. J. Med. Internet Res. **21**, 4 (2019). https://doi.org/10.2196/11029
15. Concaro, S., Sacchi, L., Cerra, C., Fratino, P., Bellazzi, R.: Mining health care administrative data with temporal association rules on hybrid events. Methods Inf. Med. **50**(2), 166–179 (2011). https://doi.org/10.3414/ME10-01-0036
16. Concaro, S., Sacchi, L., Cerra, C., Fratino, P., Bellazzi, R.: Mining healthcare data with temporal association rules: Improvements and assessment for a practical use. In: Combi, C., Shahar, Y., Abu-Hanna, A. (eds.) AIME 2009. LNCS (LNAI), vol. 5651, pp. 16–25. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02976-9_3
17. Eisenstadt, A., Liverpool, S., Metaxa, A.M., Ciuvat, R.M., Carlsson, C.: Acceptability, engagement, and exploratory outcomes of an emotional well-being app: mixed methods preliminary evaluation and descriptive analysis. JMIR Format. Res. **5**, 11 (2021). https://doi.org/10.2196/31064
18. Go, E., Lee, S., Yoon, T.: Analysis of ebolavirus with decision tree and Apriori algorithm. Int. J. Mach. Learn. Comput. **4**, 6 (2014)
19. Han, J., Pei, J., Tong, H.: Data mining: Concepts and Techniques. Morgan Kaufmann, Los Altos (2011)
20. van Heerden, A., Tomlinson, M., Swartz, L.: Point of care in your pocket: a research agenda for the field of m-health. Bull. World Health Organ. **90**(5), 393–394 (2012). https://doi.org/10.2471/BLT.11.099788
21. Huberty, J., Green, J., Glissmann, C., Larkey, L., Puzia, M., Lee, C.: Efficacy of the mindfulness meditation mobile app "calm" to reduce stress among college students: randomized controlled trial. JMIR Mhealth Uhealth **7**, 6 (2019). https://doi.org/10.2196/14273
22. Moodie Inc. 2021. Feeling Moodie. https://feelingmoodie.com/#home. Accessed 4 Jan 2022
23. Katragadda, S., et al.: Association mining based approach to analyze COVID-19 response and case growth in the United States. Sci. Rep. **11**(1), 1–12 (2021). https://doi.org/10.1038/s41598-021-96912-5
24. Kumbhare, T.A., Chobe, S.V.: An overview of association rule mining algorithms. Int. J. Comput. Sci. Inf. Technol. **5**(5), 927–930 (2014). ISSN 0975-9646
25. Lee, D.G., Ryu, K.S., Bashir, M., Bae, J.W., Ryu, K.H.: Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. J. Med. Syst. **37**(2), 1–10 (2013). https://doi.org/10.1007/S10916-012-9896-1/TABLES/6
26. Li, T., Zhang, M., Cao, H., Li, Y., Tarkoma, S., Hui, P.: What apps did you use?: Understanding the long-term evolution of mobile app usage. In: The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020, pp. 66–76 (2020). https://doi.org/10.1145/3366423.3380095
27. Manimaran, J., Velmurugan, T.: A survey of association rule mining in text applications. In: 2013 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2013. IEEE (2013). https://doi.org/10.1109/ICCIC.2013.6724258

28. Maquee, A., Shojaie, A.A., Mosaddar, D.: Clustering and association rules in analyzing the efficiency of maintenance system of an urban bus network. Int. J. Syst. Assur. Eng. Manage. **3**(3), 175–183 (2012). https://doi.org/10.1007/S13198-012-0121-X/TABLES/7
29. Murray, A., Lyle, J.: Patient adoption of mHealth. IMS Institute for Healthcare Informatics, September, pp. 1–63 (2015). www.theimsinstitute.org
30. Nahar, J., Tickle, K., Shawkat, A., Chen, Y.-P.: Diagnosis heart disease using an association rule discovery approach. In: Proceeding of the IASTED International Conference (2009)
31. Ordonez, C., Ezquerra, N., Santana, C.A.: Constraining and summarizing association rules in medical data. Knowl. Inf. Syst. **9**(3), 259–283 (2006). https://doi.org/10.1007/S10115-005-0226-5
32. Pan, H., Li, J., Wei, Z.: Mining interesting association rules in medical images. In: Li, X., Wang, S., Dong, Z.Y. (eds.) ADMA 2005. LNCS (LNAI), vol. 3584, pp. 598–609. Springer, Heidelberg (2005). https://doi.org/10.1007/11527503_71
33. Ribeiro, M.X., Bugatti, P.H., Traina, C., Marques, P.M.A., Rosa, N.A., Traina, A.J.M.: Supporting content-based image retrieval and computer-aided diagnosis systems with association rule-based techniques. Data Knowl. Eng. **68**(12), 1370–1382 (2009). https://doi.org/10.1016/J.DATAK.2009.07.002
34. Ribeiro, M.X., Traina, A.J.M., Traina, C., Azevedo-Marques, P.M.: An association rule-based method to support medical image diagnosis with efficiency. IEEE Trans. Multimed. **10**(2), 277–285 (2008). https://doi.org/10.1109/TMM.2007.911837
35. Sharma, N., Verma, C.K.: Association rule mining: an overview, vol. 5, pp. 10–15 (2014)
36. Tandan, M., Acharya, Y., Pokharel, S., Timilsina, M.: Discovering symptom patterns of COVID-19 patients using association rule mining. Comput. Biol. Med. **131** (2021). https://doi.org/10.1016/J.COMPBIOMED.2021.104249
37. Thangam, M., Vanniappan, B.: Mining association rules in dengue gene sequence with latent periodicity. Comput. Biol. J. **2015**, 1–10 (2015). https://doi.org/10.1155/2015/839692
38. Wang, X., Smith, M.R., Rangayyan, R.M.: Mammographic information analysis through association-rule mining. In: Canadian Conference on Electrical and Computer Engineering, vol. 3, pp. 1495–1498 (2004). https://doi.org/10.1109/CCECE.2004.1349689
39. Zaïane, O., Antonie, M.: Classifying text documents by associating terms with text categories. In: ADC 2002: Proceedings of the 13th Australasian Database Conference, pp. 215–222 (2002)
40. The Personal Information Protection and Electronic Documents Act (PIPEDA) - Office of the Privacy Commissioner of Canada. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/. Accessed 9 Jan 2022

Application of Shapley Additive Explanation Towards Determining Personalized Triage from Health Checkup Data

Luo Sixian(✉), Yosuke Imamura, and Ashir Ahmed

Faculty of Information Science and Electrical Engineering, Kyushu University,
Fukuoka, Japan
`luo.sixian.535@s.kyushu-u.ac.jp`
`http://socialtech.gramweb.net`

Abstract. Machine learning has become a powerful tool to assist humans making decisions. In most cases, machine learning models act like a black box, a user can only view the outcome without knowing the decision-making process or the deciding factors. Explainable AI has shown good performance in interpreting prediction models and identifying the influential parameters behind the prediction/decision. Our previous works have been analyzing health checkup data collected by a digital healthcare system, called Portable Health Clinic (PHC), developed by us. The system uses a standard logic set based on WHO recommendations to triage the health status of a patient. The triage used in PHC is almost a static standard logic set that works for any patient at any age. We argue that the triage logic should vary from person to person. This paper attempts to use explainable AI to check whether triage could be personalized. An experiment has been carried out over a health check-up data set (N = 44,460), by applying XGBoost, a popular machine learning algorithm to predict a patient's health status (risky or not risky). An eXplainable AI (XAI) technique called SHAP is used to explain the prediction results. The SHAP value clearly indicates that each health parameter (BMI, Blood Pressure, hemoglobin, etc.) has different cut-off points for different age groups, which suggests that the threshold to determine one's health status is different and can be obtained. The results will be useful to improve the existing triage static logic. This paper demonstrates cut-off points for BMI and Blood Pressure (Systolic) for two age groups which is an indication of group triage. Our future work will search for the individual cut-off point for developing personalized triage. The obtained cut-off points need to be verified by health professionals.

Keywords: Explainable AI · Health Checkup · Personalized Triage

1 Introduction

Artificial intelligence (AI) and machine learning(as subset of AI) have shown an extremely high performance for many applications in various fields of science and

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 496–509, 2023.
https://doi.org/10.1007/978-3-031-34586-9_33

technology. It allows a system to make predictions from existing data which can enable learning, reasoning, and decision-making. Machine learning algorithms can handle millions of data sets in a few seconds while the understanding of what happens in the models is not advancing at the same pace. The issue is more serious in sophisticated models such as deep neural networks [1,4]. Explainable AI (xAI) emerged to make it possible to build a model which can be interpreted and understood by the users or developers.

Explainable AI is a generic term to describe artificial intelligence. The results of the solution can be understood by humans. By applying the xAI techniques to the model which handles the data, the explanation for the results can be obtained. One way to achieve explainable AI is to develop powerful and fully explainable models, such as deep k-nearest neighbors and teaching explanations for decisions [1]. Another way is to explain the output of well-established machine learning models, instead of replacing models. An example is the Local Interpretable Model-agnostic Explanations (LIME) developed by Ribeiro et al. [2]. LIME is a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner by learning an interpretable model locally around the prediction.

The most common explanations for classification models are feature importance [3]. Feature importance refers to the usefulness of the features to make certain decisions. More specifically, it describes the individual contribution of the corresponding feature to predict a target variable. Generally, feature importance can be divided into modular global and local importance. Global importance measures the importance of the whole model while a local importance measures the importance of one observation. One example of global feature importance is the calculations of Gini impurity in decision tree. Decision tree is a tree-like model which can cope with classification problems by starting at the root of a tree and taking the branch appropriate to the outcome until encountering a leaf node. Gini impurity will be calculated in each node which can measure the feature importance of the entire data set. Based on the B-Logic and the values of each checkup item, health status/risk level as a new variable for every patient can be determined.

The effectiveness of the Explainable AI technique has been examined in many research activities. Authors in [1] assessed the LIME, an xAI framework on a tabular dataset to predict rain, and it has been shown that LIME helps to increase model interpretability. Furthermore, Nohara et al. [4] adopted Shapley additive explanation (SHAP) to interpret a gradient-boosting decision tree model using real hospital data. In their experiments, the authors found that the interpretation by SHAP was mostly consistent with that of the existing methods. Another approach based on the XGBoost algorithm and SHAP method was proposed in [5]. It showed the ability to provide meaningful explanations of the prediction model by identifying the influential risk factors. These works are useful for interpreting machine learning models and can uncover the underlying relationships between features and outcomes.

1.1 Portable Health Clinic System

Portable Health Clinic (PHC) is a digital primary health screening system that aims to control non-communicable diseases (NCDs) and to provide affordable primary healthcare services to general people [6–12]. A PHC box contains basic diagnostic tools which can be easily carried by a female health worker, shown in Fig. 1.

A health worker visits a patient's home with the PHC box, takes the clinical measurements, and uploads the health-related information to the online PHC database server. This data can be accessed by remote doctors or researchers as shown in Fig. 2 [13]. In order to improve the quality of PHC service and make proper decisions, an understanding of the PHC big data is indispensable. xAI can be used to compute the explanations for any black-box model with high accuracy. By the interpretable understanding of the machine learning models which processed PHC big data, PHC should ideally provide actionable and helpful advice for patients' prevention.

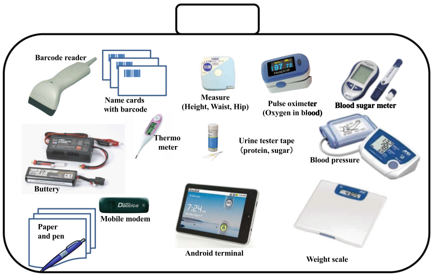


Fig. 1. A prototype of the PHC with 12 basic diagnostic tools which can measure more than 12 clinical parameters [7].

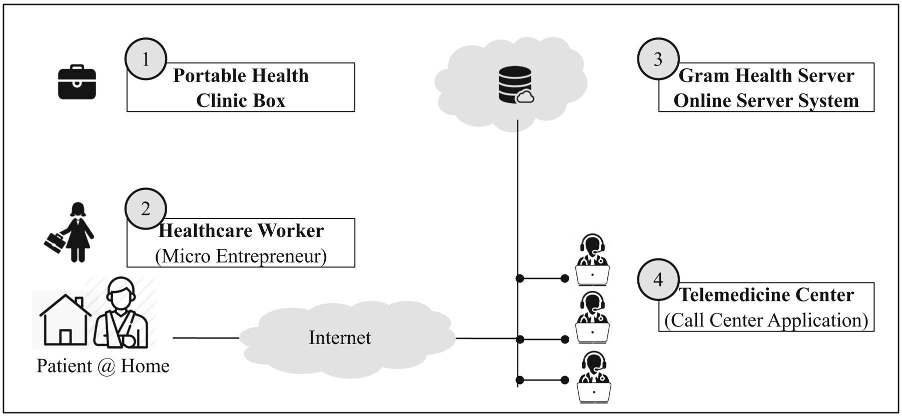


Fig. 2. PHC System Architecture. The system comprises four components: (1) PHC Device Box (2) Health Worker (3) Database System and (4) Telemedicine Center [14].

1.2 Research Motivation and Objective

In the PHC system, a standard triage logic(we call it B-Logic or Bangladesh Logic) is used to judge the health status of a patient. The results are graded in four risk categories: green (healthy), yellow (caution), orange (affected), and red (emergency) as shown in Fig. 3 [7,14]. Based on the values of each checkup item in B-Logic, the health status level for each patient can be added as a new variable.

By design, the B-Logic is almost static for all the patients. In B-logic (as in Fig. 3), the gender of a patient is the only dynamic variable for three parameters e.g. Waist, Waist-Hip Ratio, and Blood Uric Acid. These can be considered as a group behavior of the logic. However, health behavior should vary from person to person. Therefore, a personalized triage needs to be designed. However, it is not an easy task for a human being manually design and follow such a personalized triage. Doctors do it from their own experience. A machine learning tool can learn from previous experience to determine a personalized triage to make a personalized clinical decision. Thus, it is important to develop a personalized logic that can determine the health status of an individual patient. The objective of this paper is to investigate whether personalized triage can be obtained by interpreting the explainable AI tools.

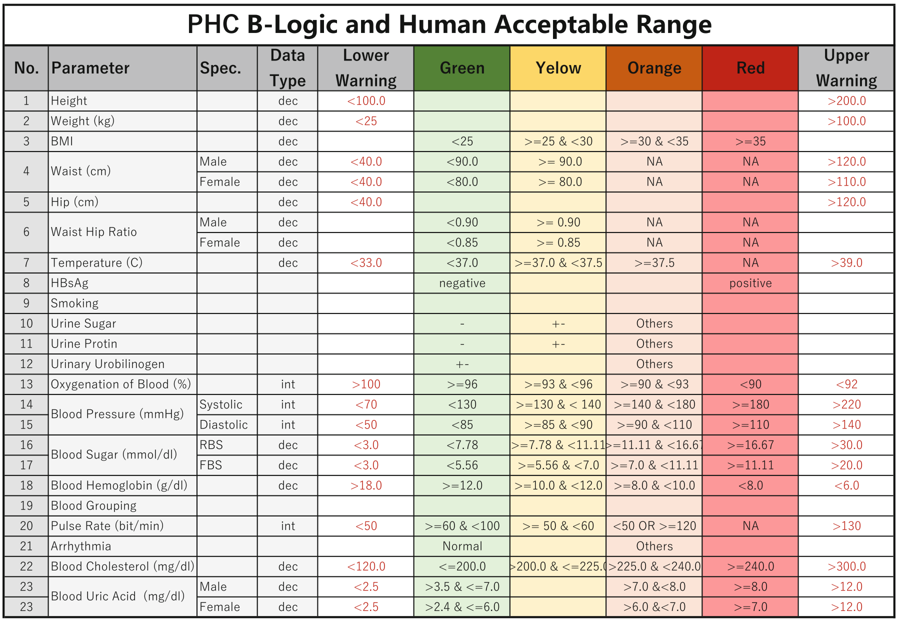
The rest of the paper is organized as follows. Section 2 introduces the techniques and analysis framework used in this research. Section 3 discusses the process of the experiment and the findings. Finally, the conclusions and future research directions are presented in Sect. 4.

2 Prediction and eXplainable AI

This section describes the technologies that are used for prediction and their explanations. A machine learning algorithm named XGBoost is used to predict the health status from health checkup data. A popular explainable AI framework, SHAP (SHapley Additive exPlanations) is used to explain the predicted health status.

2.1 XGBoost

Among many machine learning algorithms for prediction, XGBoost (eXtreme Gradient Boosting), has gained significant attention in the last few years for its high performance in solving data science problems. XGBoost is an implementation of gradient boosted decision tree algorithm proposed by Chen et al. [17]. It can construct boosted trees efficiently, and operate in a parallel way, to solve classification, regression, and ranking problems [18]. This work selected XGBoost to predict the health status of the PHC patients. The model training and data analysis were performed with Python 3.7, using the XGBoost package and scikit-learn library.



PHC B-Logic and Human Acceptable Range

No.	Parameter	Spec.	Data Type	Lower Warning	Green	Yelow	Orange	Red	Upper Warning
1	Height		dec	<100.0					>200.0
2	Weight (kg)		dec	<25					>100.0
3	BMI		dec		<25	>=25 & <30	>=30 & <35	>=35	
4	Waist (cm)	Male	dec	<40.0	<90.0	>= 90.0	NA	NA	>120.0
		Female	dec	<40.0	<80.0	>= 80.0	NA	NA	>110.0
5	Hip (cm)		dec	<40.0					>120.0
6	Waist Hip Ratio	Male	dec		<0.90	>= 0.90	NA	NA	
		Female	dec		<0.85	>= 0.85	NA	NA	
7	Temperature (C)		dec	<33.0	<37.0	>=37.0 & <37.5	>=37.5	NA	>39.0
8	HBsAg				negative			positive	
9	Smoking								
10	Urine Sugar				-	+-	Others		
11	Urine Protin				-	+-	Others		
12	Urinary Urobilinogen				+-		Others		
13	Oxygenation of Blood (%)		int	>100	>=96	>=93 & <96	>=90 & <93	<90	<92
14	Blood Pressure (mmHg)	Systolic	int	<70	<130	>=130 & < 140	>=140 & <180	>=180	>220
15		Diastolic	int	<50	<85	>=85 & <90	>=90 & <110	>=110	>140
16	Blood Sugar (mmol/dl)	RBS	dec	<3.0	<7.78	>=7.78 & <11.11	>=11.11 & <16.6	>=16.67	>30.0
17		FBS	dec	<3.0	<5.56	>=5.56 & <7.0	>=7.0 & <11.11	>=11.11	>20.0
18	Blood Hemoglobin (g/dl)		dec	>18.0	>=12.0	>=10.0 & <12.0	>=8.0 & <10.0	<8.0	<6.0
19	Blood Grouping								
20	Pulse Rate (bit/min)		int	<50	>=60 & <100	>= 50 & <60	<50 OR >=120	NA	>130
21	Arrhythmia				Normal		Others		
22	Blood Cholesterol (mg/dl)		dec	<120.0	<=200.0	>200.0 & <=225.	>225.0 & <240.0	>=240.0	>300.0
23	Blood Uric Acid (mg/dl)	Male	dec	<2.5	>3.5 & <=7.0		>7.0 &<8.0	>=8.0	>12.0
23		Female	dec	<2.5	>2.4 & <=6.0		>6.0 &<7.0	>=7.0	>12.0

Fig. 3. Concept of B-Logic and Human Acceptance Range for each clinical parameter. Each health checkup item is compared against a risk stratification matrix based on International diagnosis standards (WHO) [6].

2.2 Shapley Additive Explanation

There are various xAI techniques such as LIME (Local Interpretable Model-Agnostic Explanations) [2], SHAP (SHapley Additive exPlanations) [16], GRAD-CAM (GRADient Class Activation Mapping) [19], DeepLIFT(Deep Learning Important FeaTures) and so on [20]. Among them, SHAP has been proven to show a powerful and insightful measure of the importance of a feature in machine learning model [15]. Therefore, we applied it in this work to explain the prediction results by XGBoost.

As mentioned in Sect. 1.1, tree-based machine learning algorithms like decision tree or XGBoost can provide interpretation by calculating the gini impurity or gain, which quantified the contribution to the outcome by each feature. However, Lundberg et al. found that gain is inconsistent and proposed a unified framework for interpreting predictions, SHAP [4,16]. It is a game theoretic approach that can explain the output of any machine learning model by applying SHAP values to represent the feature importance. SHAP values have proved to be consistent and SHAP summary plots are very useful for overviewing the results.

2.3 Towards Developing a Personalized Logic/Triage

The final goal of this research is to investigate the possibility of identifying personalized triage by analyzing past health records by using an eXplainable AI tool. A personalized logic/triage is a cut-off point to categorize the person to be healthy or not healthy. The current NCD triage developed by WHO for Bangladesh and applied by the PHC system is a generalized one for all the population. This paper argues that it can be made more appropriate by developing a personalized one. As for the first step towards that goal, this paper attempts to develop a logic for a group. By understanding how the clinical parameters contribute to the patient's prediction, it will be possible to discover the threshold to determine a patient's health status, which will lead to the development of a personalized triage logic. This section explains the process to design a personal logic and a group logic.

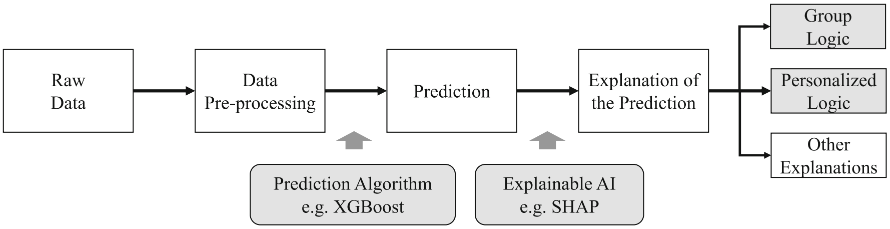


Fig. 4. A schematic view to show the development of personalized & group logic process.

A. Personalized Triage/Logic. Figure 4 shows the process to find out personalized and group logic from raw data sets. Firstly, the collected PHC data set is processed based on the criteria described in Sect. 3.1. Then the processed data is fed to the XGBoost model to predict the health status of all the patients. SHAP, an Explainable AI technique is applied to the prediction results to obtain the explanation. The SHAP value explains how the health parameters contribute to a patient's health status. The SHAP value also indicates a cut-off point between the healthy and unhealthy status. A personalized logic set can be determined from these cut-off points for each health parameter.

B. Group Logic. Figure 5 shows the process to obtain group logic based on the PHC data set. The general idea is to select the patients who have the same feature values. For example, in this work, we divided the patients into several groups based only on their age. Then the explainable AI algorithms were used to explain the models which handle different age groups. Through the explanation, we will be able to understand how the features influence the health status of the group with a certain age. The other age groups are treated in the same way. Finally, we can develop the logic to determine the health status of different age groups.

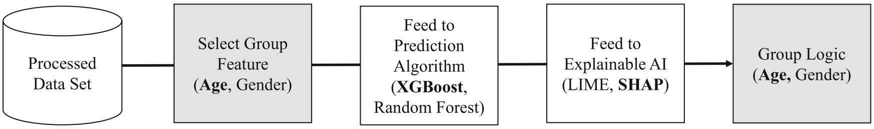

Processed Data Set
Select Group Feature (**Age**, Gender)
Feed to Prediction Algorithm (**XGBoost**, Random Forest)
Feed to Explainable AI (LIME, **SHAP**)
Group Logic (**Age,** Gender)

Fig. 5. Block diagram to present intuition for the way to achieve group logic.

3 Experiment, Results, and Discussions

The section describes the experiment environment and characteristics of the data set. The data preprocessing criteria, selection of prediction algorithm, and xAI tool are also explained.

3.1 Data Preprocessing

A list of 44,460 health checkup data have been collected since 2010 by using PHC as shown in Fig. 2. The data collection environment is explained in [10]. We have applied the data preprocessing steps (as in Fig. 6) on the raw data to solve the missing values and outlier problems [22,23]. The number of features in the raw data (V0) is 34, since it contains personal information such as name, checkup ID, and mobile number which has no relationship with our study, 7 of them will be removed. Then we applied the human acceptable range as a standard to replace the outliers with null values. Finally, after removing the rows containing the null values, we will get a complete dataset (V3). Moreover, among the 22 columns in V3, 6 unnecessary features (e.g. Checkup date, Site id) besides the target feature were removed. The final dataset (V3) contains 3085 records and 16 features that had been used for our research.

The categorical data include urinary glucose, urinary protein, urinary Urobilinogen, arrhythmia, and health status. These are encoded by label according to the triage status. For convenience, patients under Green and Yellow are considered as a new group, “Healthy”. In the same way, those who are under Orange and Red are classified as “Unhealthy”.

3.2 Prediction of Health Status by XGBoost

After the data preprocessing steps, XGBoost was applied to predict the health status of the patient with a training to test ratio of 70:30. A 10-fold cross-validation process was used to evaluate the performance of the model. The accuracy (according to Eq. 1) was 99%, which indicated that XGBoost achieved a good prediction accuracy. Furthermore, the confusion matrix is given in Table 1. Balanced accuracy is 99% which can be calculated through Eq. 2, 3, 4, where TP is true positive, FP is false positive, FN is false negative and TN is true negative.

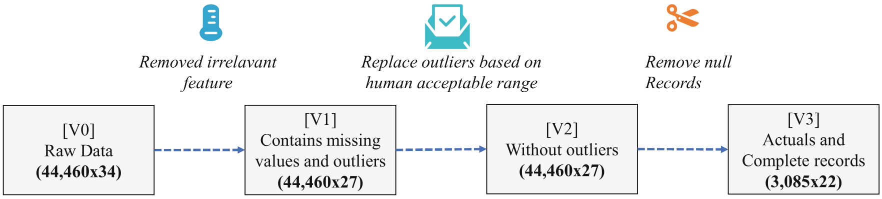

Removed irrelavant feature
Replace outliers based on human acceptable range
Remove null Records
[V0] Raw Data (44,460x34)
[V1] Contains missing values and outliers (44,460x27)
[V2] Without outliers (44,460x27)
[V3] Actuals and Complete records (3,085x22)

Fig. 6. PHC data preprocessing for prediction

Table 1. The confusion matrix for PHC data health status prediction model.

Actual	Predicted	
	Negative	Positive
Negative	596	3
Positive	0	327

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Balanced\ Accuracy = \frac{TPR + TNR}{2} \tag{2}$$

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

$$TNR = \frac{TN}{TN + FP} \tag{4}$$

Figure 7 shows the feature importance scores calculated through the gain method by XGBoost. We observe that among the 16 clinical parameters, urinary protein is detected as the most influential feature to determine health status. Arrhythmia, bp_dia, and blood_hemoglobin are also highly ranked features. Gender, waist, waist_hip_ratio, age_oncheckup, and oxygen_of_blood have the least effect on the prediction of health status. However, through this result, it's difficult to understand how every feature influences the target i.e. how the value of a feature bring positive or negative contributions to the prediction. Therefore, Shapley Additive Explanation will be used to discover the potential relationship behind the features.

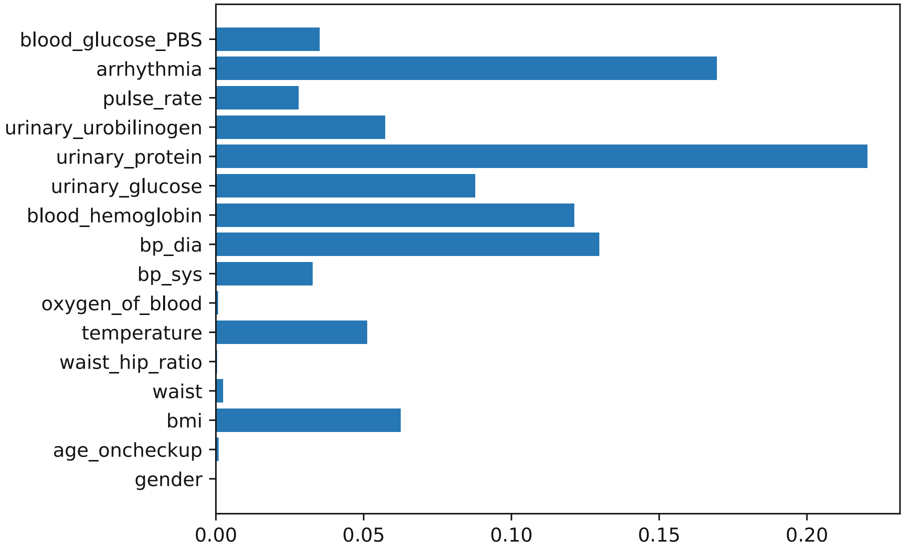


Fig. 7. The Prediction result of XGBoost. Y-axis shows different health parameters and X-axis shows the feature importance to determine health status. Global feature importance values showing urinary protein was selected as the most influential feature.

3.3 Use of SHAP to Explain the Prediction Results

The SHAP summary plot shows the positive and negative contribution of the features with the target variable. Each point on the summary plot is a Shapley value for a feature and an instance. Figure 8 shows the explanation for the XGBoost model which handled the data processed in Sect. 3.1 and it delivers the following information:

- **Feature importance**: In the y-axis direction, features are ranked in descending order according to their importance. In this study, blood hemoglobin was extracted as the most influential feature.
- **Impact**: The x-axis shows whether the feature has a positive or negative contribution to the target value. The points distributed on the positive x-axis have a positive impact on the status of unhealthy, and the negative points on the x-axis lead to a healthy status.
- **Original value**: Color indicates the value of the feature. Blue indicates low and red indicates high for that individual instance. For categorical features with only two possible values such as arrhythmia, it will take only two colors, but numerical data like bmi or blood pressure can contain the whole spectrum.
- **Correlation**: Through combining the impact and original value, we found that a low value of blood hemoglobin increases the risk, which indicates blood hemoglobin is negatively correlated with the target variable, that is, patients' health status.

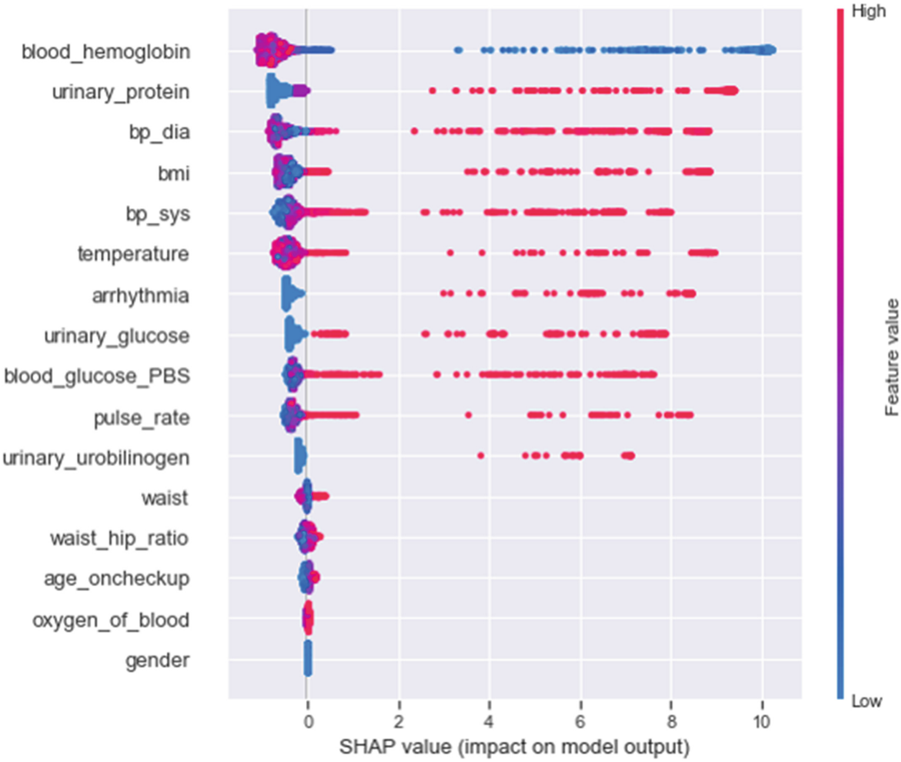


Fig. 8. SHAP Summary Plot output for explaining XGBoost. The figure shows the sorted feature importance and the relationship between each feature and target i.e. the health status. Among 16 of them, blood hemoglobin is detected as the most influential feature.

For detailed analysis, we plot the SHAP dependence plot of BMI for the different age groups in Fig. 9. It shows the relationship between BMI and its' effect on health status. Each dot is a single prediction from the dataset and the x-axis represents the value of BMI. The figure on the left shows the result by SHAP for patients whose age is 25. We observe that a higher BMI value causes a higher risk to be unhealthy. Moreover, SHAP values are above zero when BMI is higher than 26, in contrast, there are some negative dots for the 35-year-old group on the right side until a BMI value of 27.5. The results suggest that for different age groups, the standard to determine a patient's health status should be different.

Figure 10 shows the impact of blood pressure (systolic) on health status. The figure indicates that a value of approximately 130 is the threshold of a healthy status for patients whose age is 25 and the threshold for 35-year-old group patients should be about 135. Furthermore, the summary of several considered features' cut-off points is shown in Table 2. In this way, we are able to design a new standard for every feature in the PHC dataset based on patients' age.

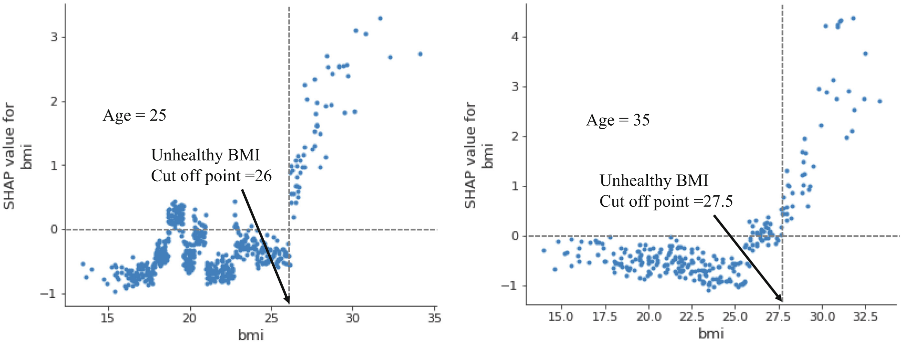


Fig. 9. SHAP Dependence Plot of BMI. The left figure shows the SHAP value for Age = 25 and the right figure shows the same for Age = 35. Two dotted lines are drawn in a way such that no negative SHAP value exist on the right side. This way, the cut-off points for both the ages are determined.

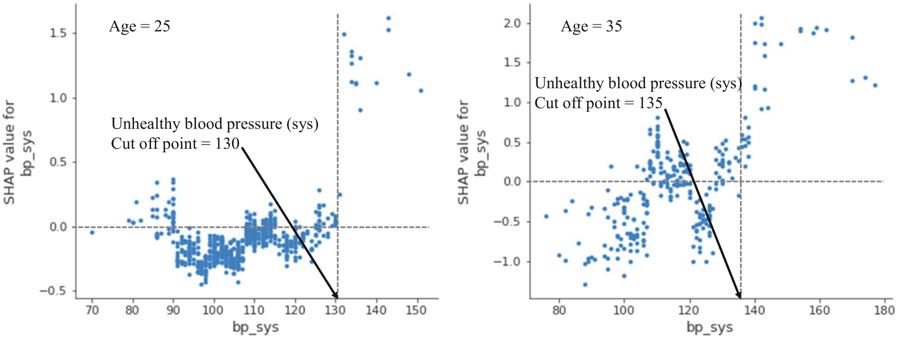


Fig. 10. SHAP Dependence Plot of blood pressure (systolic). The left side is the result of the 25-year-old group, the right side is for the 35-year-old group. The cut-off point for the age group 25 is indicated as 130, and for the age group 35, the cut-off point is 135.

Table 2. The cut-off points for different clinic parameters show the threshold of a healthy status for the two age group patients.

Feature	Cut-off Point(Age25)	Cut-off Point(Age35)
bmi	26	27.5
bp_sys	130	135
waist_hip_ratio	0.93	0.91
temperature	98.7	98.7
bp_dia	88	89
blood_hemoglobin	9.8	9.9
blood_glucose_PBS	133	141

4 Conclusions and Future Work

This paper used explainable AI to determine a cut-off point for a person's binary health status i.e. personalized triage. The research began with determining triage for an age group. A PHC data set (N = 44,460) and a popular machine learning algorithm, XGBoost were used to predict a patient's health status (risky or not risky). An eXplainable AI (XAI) technique called SHAP is used to explain the prediction results. The SHAP value clearly indicated the cut-off point for each health parameter (BMI, Blood Pressure, hemoglobin, urinary protein, etc.) for different age groups. The results suggest that the threshold to determine one's health status is different and can be obtained, which is useful for us to refine the existing triage static logic. Cut-off points for BMI were found for different age groups. For example, the cut-off point of BMI for the age group 25 was 26 whereas, for the age group 35, the value was 27.5 respectively. The obtained cut-off points need to be verified by health professionals. Moreover, the SHAP summary plot showed the ability to identify the risky feature for the PHC dataset by calculating the SHAP value as a global feature importance score. As a result, blood hemoglobin is detected as the most influential feature to determine a patient's health status and it shows that lower blood hemoglobin causes a higher risk of unhealthy.

In terms of future work, we will continue the research about Explainable AI and try to implement the novel explainable framework on PHC dataset. Through the understanding of how the clinical parameters contribute to an individual patient's prediction, we should ideally discover the threshold to determine a patient's health status, which leads to the development of a personalized triage logic. We aim to implement the SHAP dependence plot for each age group in the PHC database. The current size of the dataset for some age groups is not sufficient to build the model, imputation of the raw PHC dataset will increase the size of the data and will be considered.

Acknowledgment. The authors are grateful to the members of the Global Communication Center in Grameen Communications for their assistance in collecting PHC data in Bangladesh.

References

1. Dieber, J., Kirrane, S.: Why model why? Assessing the strengths and limitations of LIME. arXiv, abs/2012.00093. https://doi.org/10.48550/arxiv.2012.00093
2. Ribeiro, M., Singh, S., Guestrin, C.: Why Should I Trust You?: Explaining the Predictions of Any Classifier, pp. 97–101. https://doi.org/10.18653/v1/N16-3020
3. Saarela, M., Jauhiainen, S.: Comparison of feature importance measures as explanations for classification models. SN Appl. Sci. **3**(2), 1–12 (2021). https://doi.org/10.1007/s42452-021-04148-9
4. Nohara, Y., Matsumoto, K., Soejima, H., Nakashima, N.: Explanation of machine learning models using Shapley additive explanation and application for real data in hospital. Comput. Methods Programs Biomed. **214**, 106584. https://doi.org/10.1016/j.cmpb.2021.106584

5. Athanasiou, M., Sfrintzeri, K., Zarkogianni, K., Thanopoulou, A.C., Nikita, K.S.: An explainable XGBoost-based approach towards assessing the risk of cardiovascular disease in patients with Type 2 Diabetes Mellitus. In: IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 859–864 (2020). https://doi.org/10.1109/BIBE50027.2020.00146
6. Ahmed, A., et al.: Portable health clinic: a telehealthcare system for unreached communities. In: Lin, Y.-L., Kyung, C.-M., Yasuura, H., Liu, Y. (eds.) Smart Sensors and Systems, pp. 447–467. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-14711-6_18
7. Ahmed, A., Inoue, S., Kai, E., Nakashima, N., Nohara, Y.: Portable health clinic: a pervasive way to serve the unreached community for preventive healthcare. In: Streitz, N., Stephanidis, C. (eds.) DAPI 2013. LNCS, vol. 8028, pp. 265–274. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39351-8_29
8. Nohara, Y., et al.: Health checkup and telemedical intervention program for preventive medicine in developing countries: verification study. J. Med. Internet Res. **17**(1) (2015)
9. Islam-Maruf, R., Ahmed, A., et al.: Portable health clinic as a telemedicine system with appropriate technologies for unreached communities. In: Maeder, A.J., Higa, C., van den Berg, M.E.L., Gough, C. (eds.) Telehealth Innovations in Remote Healthcare Services Delivery - Global Telehealth 2020. Studies in Health Technology and Informatics, vol. 277, pp. 57–67. IOS Press BV. https://doi.org/10.3233/SHTI210028
10. Ahmed, A., Hasan, M., Sampa, M.B., Hossein, K.M., Nohara, Y., Nakashima, N.: Portable health clinic: concept, design, implementation and challenges. In: Mobile Technologies for Delivering Healthcare in Remote, Rural or Developing Regions, pp. 105–121. Institution of Engineering and Technology
11. Islam, R., Nohara, Y., Rahman, M.J., Sultana, N., Ahmed, A., Nakashima, N.: Portable health clinic: an advanced tele-healthcare system for unreached communities. Stud. Health Technol. Inform. **264**, 616–619 (2019)
12. Kikuchi, K., et al.: Portable health clinic for sustainable care of mothers and newborns in rural Bangladesh. Comput. Methods Progr. Biomed. **207**, 106156 (2021)
13. Tabassum, S., Sampa, M., Islam, R., Yokota, F., Nakashima, N., Ahmed, A.: A data enhancement approach to improve machine learning performance for predicting health status using remote healthcare data. In: 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), pp. 308–312. IEEE (2020)
14. Tabassum, S., Sampa, M., Maruf, R., Yokota, F., Nakashima, N., Ahmed, A.: An analysis on remote healthcare data for future health risk prediction to reduce health management cost. In: 11th Biennial Conference of the Asia-Pacific Association for Medical Informatics, APAMI 2020, pp. 115–119 (2020)
15. Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A.: Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accid. Anal. Prev. **136**, 105405. https://doi.org/10.1016/j.aap.2019.105405
16. Scott, M.L., Su-In, L.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, pp. 4768–4777 (2017)
17. Tianqi, C., Carlos, G.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), pp. 785–794. Association for Computing Machinery, New York (2016). https://doi.org/10.1145/2939672.2939785

18. Yang, C., Chen, M., Yuan, Q.: The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: an exploratory analysis. Accid. Anal. Prev. **158**, 106153
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision, pp. 618–626 (2017)
20. Machlev, R., Heistrene, L., Perl, M., et al.: Explainable Artificial Intelligence (XAI) techniques for energy and power systems: review, challenges and opportunities. Energy and AI 100169 (2022)
21. Hossain, N., Sampa, M.B., Yokota, F., Fukuda, A., Ahmed, A.: Factors affecting rural patients' primary compliance with e-prescription: a developing country perspective. Telemed J. e-Health **25**(5), 391–398 (2019). Epub 8 June 2018. PMID: 29882727; PMCID: PMC6534088. https://doi.org/10.1089/tmj.2018.0081
22. Imamura, Y., Abedin, N., Sixian, L., Tabassum, S., Ahmed, A.: Missing value imputation for remote healthcare data: a case study of portable health clinic system. In: The 9th International Japan-Africa Conference on Electronics, Communications, and Computations (JACECC), pp. 85–88. IEEE (2021)
23. Tabassum, S., Abedin, N., Maruf, R.I., Ahmed, M.T., Ahmed, A.: Improving health status prediction by applying appropriate missing value imputation technique. In: 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech), pp. 345–348 (2022). https://doi.org/10.1109/LifeTech53646.2022.9754794

Human-Centred Design of Pervasive Health Solutions

Dance Mat Fun - A Participatory Design of Exergames for Children with Disabilities

Gonçalo Roxo[1(✉)], Rui Nóbrega[1], and Rui Neves Madeira[1,2]

[1] NOVA LINCS, NOVA School of Science and Technology,
NOVA University of Lisbon, Lisbon, Portugal
`g.roxo@campus.fct.unl.pt`, `rui.nobrega@fct.unl.pt`
[2] Sustain.RD, Escola Superior de Tecnologia de Setúbal,
Instituto Politécnico de Setúbal, Setúbal, Portugal
`rui.madeira@estsetubal.ips.pt`

Abstract. Physiotherapy can lead to a long and tedious process where the progress of rehabilitation is usually not immediately visible. Exergames introduce a fun factor to the therapy while keeping rehabilitation as the primary goal, promoting patient engagement to foster adherence to therapy. This paper presents a participatory design of exergames for children with disabilities. We intended to explore the interaction with dance mats as a means of promoting rehabilitation and therapy adherence using a fun and low-cost off-the-shelf device for interaction. The games are designed to be adaptable, with different degrees of speed, difficulty and modality. Using the mat, the children can play standing up, seated or laying down. We worked with a rehabilitation clinic center for children to create a suite of exergames following a design thinking methodology, benefiting from the collaboration between developers, therapists, and patients. Therapists and children participated actively in the process design to help us create games suited to their needs. We present the participatory design process with focus on the user study that provides important insights on what works for certain groups of children with disabilities.

Keywords: Exergames · Physiotherapy · Children with special needs · Participatory design · Low-cost controllers · Dance mat

1 Introduction

Traditional therapy must be repeated for long periods, leading to patients that may become demotivated and lose focus on the therapy program [9]. Therapists are always in need of tools that can help them mask this repetitiveness, making it important to research ways of keeping the patients motivated and engaged in

---

This work was supported by NOVA.ID.FCT/NOVA LINCS (UIDB/04516/2020).

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 513–527, 2023.
https://doi.org/10.1007/978-3-031-34586-9_34

the therapy. It is known that motivated patients spend more time and effort into promoting their recovery [14].

By combining video games with physiotherapy, therapy sessions can be more motivating [12] and accessible without a significant increase in healthcare costs. Al-Nasri and Salim [1] describe two main issues with physiotherapy adherence: (1) the barriers patients face that lead to non-adherence, from low levels of physical activity before physiotherapy to depression and poor social support; (2) the lack of adoption of new technologies that could encourage intrinsic motivation for adherence in patients. Al-Nasri and Salim [1] claim the lack of adoption of new technologies for physiotherapy is more apparent than in other health sectors since the nature of rehabilitation requires hands-on applications, where clinicians need to be able to control parameters such as speed, duration, and difficulty.

Creating games for children with disabilities may generate particular challenges that are easier to address by employing a participatory design approach. Since typical interaction designers do not usually have a deep understanding of the children's needs, the process should always be multidisciplinary, where experts in the rehabilitation area should be involved, as much as possible, to work with the system developers to find the requirements needed and design the solution to achieve the therapy goals [10,13]. A participatory design provides three main benefits [15]: 1) a better understanding of problems, 2) the building of realistic expectations in target groups, 3) the empowerment of marginalized groups by giving them a stake in the technology design, giving them with a sense of ownership and control over shaping one's own environment.

In this paper, we present research work focused on exploring the creation of therapy-oriented games under a participatory design setting, as a research goal. We enlisted the collaboration of therapists from a clinic with deep experience and expertise on developmental disabilities - a group of lifelong conditions caused by an impairment in the physical, learning, speech, or behavior areas. Children diagnosed with such disabilities typically require therapy, or other services, to address behavioral and developmental challenges [30].

As a second research goal, the team decided to explore the use of simple, yet motivating, interaction devices like a dance mat (similar to the one from Dance Dance Revolution[1]). The mat has distinct characteristics that make its usage very flexible, as it can be displaced horizontally on the floor (Fig. 1 (left)), where it can be used with feet, hands, and even the body, or vertically on a wall (Fig. 1 (right)), to be used with the hands, for instance.

Therefore, we have applied a design thinking [8] process in which the therapists participated actively in the define, ideate and test stages. The partnership produced four exergame prototypes that can be played with the dance mat and have several adjustable parameters to be more adaptable.

The paper is organized as follows. Section 2 contextualizes the background and related work. Afterwards, we present our participatory design approach in Sect. 3. A first user study appears in Sect. 4 and results are discussed in Sect. 5. Finally, conclusions and Future Work are summarized in Sect. 6.

---

[1] https://www.konami.com/amusement/products/am_ddr/ (last access: sep 2022).

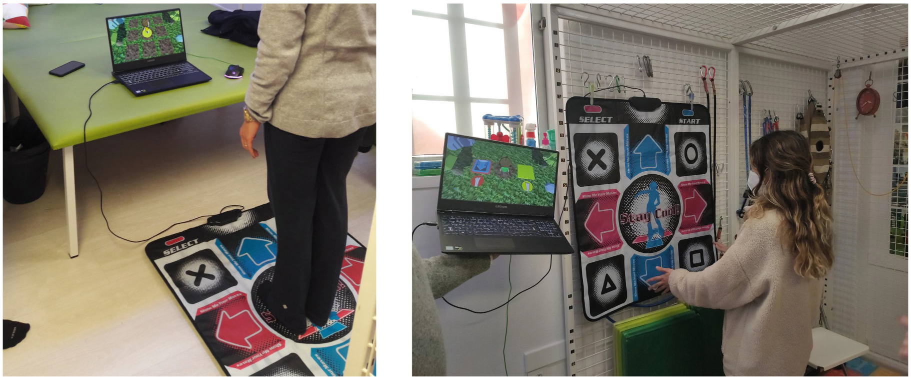


Fig. 1. Physiotherapists testing the prototypes and exploring the mat in different ways, using: (**left**) lower limbs; (**right**) upper limbs.

2 Background and Related Work

We focused mainly on the subject of therapy for children and on how others approached the task of developing serious games.

Therapy can be habilitative when the goal is to develop new skills or rehabilitative if the patient had the skills previously. Its goals are, among others, to help the patient achieve an appropriate level of functional skills for their development; lower the impact of body part impairments, and ensure that families are supplied with support and training to carry over the therapy into other settings [17].

Traditional physiotherapy programs in children with Cerebral Palsy (CP) have been shown to improve muscle strength, local muscular endurance, and overall joint range of motion. A program [25] of progressive resistive exercises is used to improve muscle strength, while a program that uses low resistance and more repetitions will enhance local muscle endurance. To maintain and improve joint mobility, patients need to repeat several passive range of motion exercises, and to prevent joint contracture, patients also need to stretch. Traditional occupational therapy has also been used to improve fine motor skills and the use of the upper limbs to increase the ability to perform daily living tasks. Occupational therapy also empowers children with the knowledge and skills they need for self-care and information processing [25].

Despite its benefits, low motivation and engagement with therapy are very common problems. Pervasive computing technologies can make play-based occupational therapy more effective by embedding digital technology into playful activity [22]. In occupational therapy, an effective means to motivate a change in a child's behavior is by designing playful activities that leverage the child's desire to play. Video games designed specifically for rehabilitation can do the work without adding significant costs to both the healthcare system and patients [5].

Moreover, the use of participatory design [4,18] is important to have a wider discussion about different aspects and perspectives on what a solution should be. The integration of experts in the design team can be used to leverage their knowledge to create a more meaningful experience to the players, as Thompson et al. [28] report in their article about how behavioral science guided the design of a serious video game to prevent Type 2 diabetes and obesity. The final solution should be the result of an iterative process involving different actors, such as, therapists, children, parents, designers and developers.

Exergames can be an important component for the rehabilitation process, but aspects, such as, Fun vs Effectiveness, Variations of Exercise, Goals, Positive Feedback and how to present Negative Outcomes should be considered [2]. Burke et al. [9] identified two main principles of game design to be very important in the development of serious games: meaningful play and challenge.

Meaningful play stems from the relationship between players' actions and the system's outcome. The actions must have feedback when performed to give players the awareness they need regarding their choices. In rehabilitation games, the feedback [29] should take into account the difficulties patients playing the game might have. Failure should be handled more conservatively with an initial focus on increasing engagement and then rewarding players with success. Correct identification of failure can also be used to teach players by making them reconsider strategies [27].

Challenge is closely related to the difficulty of the game objectives. Generally, starting with a lower level and gradually increasing throughout the game according to the player's ability. A balanced experience has shown increased motivation and higher perceived levels of fun and fairness, even when players compete against each other [19,20].

A study [11] comparing exergames designed for individuals with autism spectrum disorder (ASD) with commercially available ones, on their performance as a tool for supporting visual-motor coordination training, showed exergames designed for individuals with ASD excel in many aspects when compared to commercial games and provide an overall better experience for the players. Another study [21] developed two games using the dance mat and reported balance improvement on patients with CP.

Games for rehabilitation have to be prepared to adjust their parameters according to the patients' difficulties. After an evaluation from the therapist, the game speed could be adjusted to set its pace [6]. The position of objects could also be adjusted to account for different ranges of motion, making the game's difficulty more appropriate for a specific patient. It is always better to have a conservative starting difficulty when the user is being introduced to the system to minimize the risks of failure.

Despite their limited motor capacity, children with cerebral palsy still want to play action-oriented games similar to the ones played by their peers without motor disabilities [16]. Through a year-long participatory design process with children with CP, researchers have found that it is, indeed, possible to develop such games for children with disabilities. However, they also recognize that using what they call "traditional guidelines" can still be the correct choice for some kinds of games, such as those focusing on stretching and balancing actions where the goal is not to encourage cardio-vascular exercise. Following those guidelines will also make the user group that can benefit from the games as large as possible. The guidelines are as follow: avoid fast pace, do not require precise timing, provide a simple control scheme, do not require multiple simultaneous actions, avoid repeated inputs (button mashing), and automate the player's input.

3 Participatory Design of Rehabilitation Exergames

We employed a design thinking approach consisting of five different stages: Empathize, Define, Ideate, Prototype and Test [8]. We had the active collaboration of therapists (experts) from a clinic with deep experience working with children with different disabilities. Their integration in our design team allowed us to leverage their therapy expertise to create more meaningful experiences.

The **Empathy** stage of the process was already taken through previous collaboration on different projects, mainly "just Physio kidding" [23], which is a solution based on the use of serious games with Kinect-based natural user interfaces and personalisation. In the **Define** stage, the team focused on discussing and selecting the therapeutic exercises that needed motivational support to be executed by children, like repetitive reaching exercises or the ones that require moving to different positions to train mobility and balance while standing. It was established that therapists sometimes used commercial games in their sessions, but these were often hard to adapt to their patients' capabilities and did not allow them to collect relevant data to measure progress. Therefore, they would like to see exergames where they could still use off-the-shelf controllers but have more control over the game parameters. They considered the dance mat a great controller to explore due to its possibilities regarding different therapeutic exercises' dimensions, because of its ability to, not only, be used as a stepping mat, but also be folded or hung on a wall, for example.

Reaching the **Ideate** stage, we had brainstorming sessions with the therapists to understand how we could build games that benefit the therapy of children followed by the clinic. When possible, we also asked children what themes they enjoyed to make the games more appealing. We settled on a game suite where children would have to use body movements to control the games so it would be possible to train balance and mobility. The idea of having different games would also allow us to test which kind would be more suitable for the children at the clinic and their characteristics. If games had an element of symbol/pattern recognition, they could also be used to train focus and stimulate cognitive function. Again, we found the dance mat could satisfy these requirements while being very affordable, easy to use, and requiring no calibration, unlike other sensors like the Microsoft Kinect.

With the suggestions and ideas collected in the define and ideate steps, we initiated the development (**Prototype** stage) of four prototype games. They were called"Matchmat", "Left or Right?", "Crazy Car" and "Fishing". With the dance mat being a simple interaction device, all games follow the same interaction principles with varying degrees of complexity. We tried to integrate different

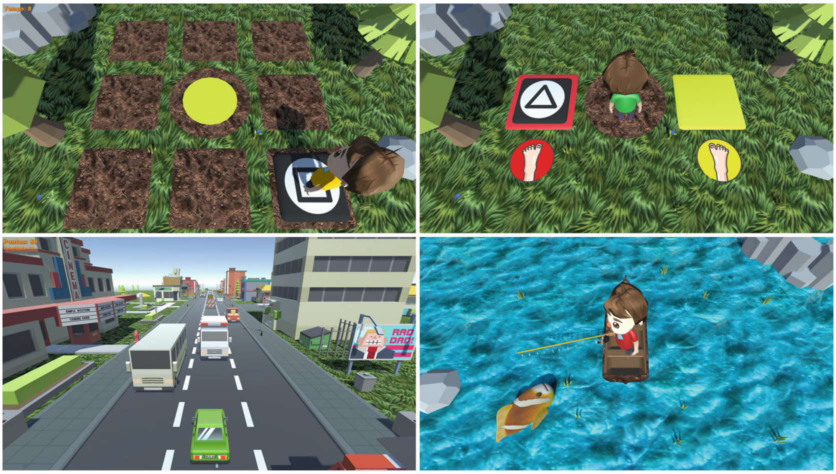


Fig. 2. **Top Left:** Depiction of *Matchmat*; **Top Right:** Depiction of *Left or Right?*; **Bottom Left:** Depiction of *Crazy Car*; **Bottom Right:** Depiction of *Fishing*

game modes and input layouts that would allow the games to be more suited to a higher number of patients depending on their physical and cognitive strengths. In every game, it is also possible to adjust parameters, such as, speed, button and obstacle placement, a.o., as required by the therapists. The games had a first iteration as proof of concept, with simple 2D visuals to allow the therapists to see the dance mat interaction in person and test it with two children. After this small test session and getting their approval, a more developed version of the games was created (Fig. 2) with updated 3D graphics, and better visual and audio feedback because the children had found the previous design boring.

Matchmat (G1): Matchmat shows a representation of the dance mat on-screen with a human character in the middle that the children can recognize as a reference. Buttons will show up on the screen for a limited time, with their virtual positioning matching the one in the real world. There are two modes in this game, one where only the correct button is displayed and another where all the buttons are always shown on the screen, with the correct one being highlighted with a red animated square. This way, children will have an easier time identifying which button to press and can train the matching between real-world positioning and the visual representation of the buttons, as well as having to use real-life movements to reach each button. ***Goal:*** Improve equilibrium, Real-virtual visual matching, Mobility.

Left or Right? (G2): Left or Right? Is a more complex variation of Matchmat. There are two colored squares on the screen with a symbol representing the left and right feet below them. When a button appears on the red square, the patient must press it using their left foot. When it appears on the yellow button, it must be pressed with the right foot. In this game, the children no longer see the whole

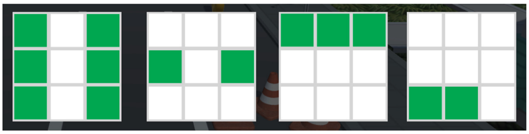


Fig. 3. Dance mat controller layout options for *Crazy Car*. Highlighted squares represent the buttons used to play the game (from left to right: V1, V2, V3, V4).

representation of the mat in the virtual space. They must not only recognize the symbol and find its position on the mat but also decide which is the correct foot to use. ***Goal:*** Improve Equilibrium, Real-virtual visual matching, Mobility, Limb laterality recognition.

Crazy Car (G3): Instead of having to match symbols shown on the screen, this game is an "endless runner" style game where a car continuously moves forward. The patients must use the buttons on the mat to dodge obstacles placed on the road. For more adaptability, this game offers four controller layouts that allow the game to be played while standing, sitting, or possibly, laying down. The control layouts (Fig. 3) use button presses to make the car switch lanes, with variations V1, V2, and V4 using the buttons on the left and right sides of their layout to move the car in the respective direction. Variation V3 matches each button from the first row of the mat to each of the three lanes the car can occupy, allowing to move instantly between lanes. ***Goal:*** Improve Equilibrium, Real-virtual visual matching, Mobility, Reflexes, Decision making.

Fishing (G4): Fishing is harder to play because it requires a higher cognitive capacity. The grid representation of the mat is not visible on the screen, and no buttons appear in this game. Instead, a fish will appear in a around the player character in a zone that represents a position on the mat. The patients must then understand in which position the fish is in relation to them without the help of any other visual cues, like they had in the previous games, and press the correct button. Once they press the correct button, a mini-game is triggered where they must counter the movement of the fish by going left or right on the mat to catch it. ***Goal:*** Improve equilibrium, Real-virtual visual matching, Mobility, Situational awareness.

4 User Study

In the **Test** stage of the design, we ended-up conducting a user study focused on testing the usability of mat-controlled games and how children with different disabilities would be able to interact with them. The therapists did a pre-selection of children they considered able to use the system at the clinic. We were able to achieve a sample size of ten participants with varied diagnostics and abilities, which allowed us to test the mat in different ways. In Table 1, we can see these characteristics which include, for the patients with cerebral palsy, the Gross

Table 1. Diagnosis and Characteristics of the patients that tested our system.

ID	Diagnosis	Characteristics
P1	Spinal Muscular Atrophy - Type I	General muscle weakness. Restricted amplitude of movement (mainly elbows, knees, and hip)
P2	Articulation and Phonological Disorders	Neurotypical development
P3	CP - Spastic Diplegia	GMFCS II. Able to walk, run, and climb stairs with the support of a railing. Difficulty jumping and balancing on one foot
P4	CP- Spastic Tetraparesis	GMFCS IV. Higher function with the upper left limb. Limited amplitude of movement of knees and hip. Able to roll independently. Slight cognitive deficit
P5	CP- Spastic Tetraparesis	GMFCS IV. Able to roll with assistance and remain seated with the support of the upper limbs. Difficulty dissociating lower limbs. Cognitive deficit
P6	CP - Spastic Diplegia	GMFCS III. Lower limb constraints and difficulty maintaining balance while standing without support. Can perform all posture transfers but with a higher risk of falling. Prefers to stand with flexed knees and hip. Slight cognitive deficit
P7	CP - Left Hemiparesis	GMFCS I. Bigger upper limb constraints and difficulty maintaining balance while standing on one foot. Able to walk, run and climb stairs without assistance
P8	Global Developmental Delay	Lack of motor coordination and planning
P9	Autism Spectrum Disorder	Difficulties with information processing and motor planning. Difficulty performing the same activity for long periods
P10	Autism Spectrum Disorder	Difficulty focusing and managing frustration. Slight trouble with motor planning

Motor Function Classification System (GMFCS) [24] level. This is a I to V scale where V means less mobility. Participant P6 was a female, and the other nine were males. The ages ranged between three and eleven with a median of 7.5. All tests were performed with parental consent. Before each session, the children were asked if they wanted to participate and the therapist assessed how comfortable they were in the process. Testing was done over the course of two weeks, and we only had limited time with each patient. Our testing sessions had to be done during small time windows that coincided with the time patients went to the clinic for their regular therapy sessions without disrupting their treatments.

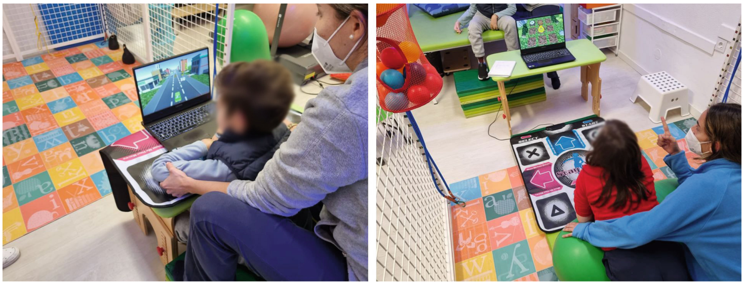


Fig. 4. Left: The mat can be folded to allow children with very limited mobility to play using elbows; **Right:** Child using Pilates ball to play a game while laying down.

Children with severely limited mobility could only play the endless-runner game while sitting down (Fig. 4, Left), supported by the therapist. The mat was placed over a table to allow them play using the elbows. Another option for children with trouble standing and maintaining balance was laying over a Pilates ball (Fig. 4, Right). That allowed them to roll with it and reach every button with the help of a therapist.

At this stage, our main concern with this first evaluation was to assess the children's acceptance of the system and the therapists' enthusiasm for using it in their therapy sessions. Due to the many different ways in which the children at the clinic were affected by their respective disabilities, the studies can not be as standardized as when working with neurotypical children with normal motor function. We decided our approach with this study would be to test the usability of mat-controlled games and how children with different symptoms would be able to interact with them. We also wanted to assess the degree of success they would have in their interactions and if we could establish some kind of patient profile that could benefit more from these games.

We found that the Fishing game was not adequate to use with the children because it was hard for them to recognize the place of the fish and press the correct button. Also, in some cases, since the character used for reference, at the center of the screen, always turns to face the fish, some children were inclined to always press the button representing the "front" position. For these reasons, we focused only on the first three games when moving forward with testing.

Therapists showed interest in the system from the start and looked forward to seeing the potential of its integration into the children's therapy sessions.

5 Results and Discussion

Given the nature of our users and their very different characteristics, we believe achieving a sample size of ten was good and allowed us to test the mat in different ways. Each child went through each game, and we logged which ones they were able to play and which variations did they play. We also registered if they needed active support to play the game and collected data, sent in real-time to an external platform [3], from each game execution.

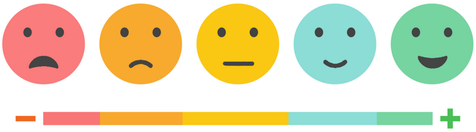


Fig. 5. Visual Likert scale used to help children answer the questions.

A questionnaire was set up to be answered by both the therapists, and, when possible, by the children after each session to assess their thoughts and suggestions about the proposed solution. The first set of questions was directed at the children, using a simplified System Usability Scale (SUS) [7] where the questions were adapted to a younger audience [26] to get their direct opinion about the games and their overall experience with the system using the support of a visual representation of the Likert scale (Fig. 5).

The second set of questions, also using SUS, was directed at the therapist that accompanied the child during the game session. The SUS questions were directed to the specific interaction of each child with the system from the therapist's point of view. Since the therapists were always involved in the development process of the system and given the many variable patient characteristics, our goal with this questionnaire was not necessarily to attain a high score with the SUS. Our interest was in using the SUS structure to understand how children with different disabilities would interact with the mat and their motivation to do so. Then, on the side of the therapists, we wanted to see if exergames designed to be controlled with a dance mat could have a role in the therapy regimes of their patients.

Regarding the first set of questions, out of the ten children, only two didn't manage to answer the questions at all. We had SUS scores ranging from 25 to 100, averaging 68.75. It makes sense that the children that were not able to play the games as intended or didn't understand them produced a lower score on the questionnaire. But we were optimistic when seeing that the children that can play gave the system very high scores. Although it is important to note that in their enthusiasm, most children tended to rate every question with the highest score, we believe that same enthusiasm shows that they were motivated to use the system and wanted to keep playing the games.

On the therapist side the system got an average SUS score of 77.5. But we would like to focus on the statement: "I think that I would like to use this system frequently with this patient." In the context of this study, the answer to this question is the most important since even when a therapist understood the system and the child had no trouble playing the games, the therapist can consider that the system does not meet the therapy goals set for a given patient, as happened with patient P2. P2 is only enrolled in speech therapy and does not need to practice the kind of movements required by our games, so their expected usage in therapy sessions was "Unlikely", as seen in Table 2, where we grouped every child based on the therapist's answer to that statement. The table also shows each child's main constraints, GMFCS level, when applicable, if they needed support to play the game, and the results for each game played.

Table 2. Expected usage of the system by each patient in their therapy sessions with their respective game results. Total error Manhattan Distance (MD) for game G1 and G2, and play time for G3.

Expected Usage	ID	Main Constraints	GMFCS	Needs Support ?	G1 MD #	G2 MD #	G3 Time (s)
Frequent	P6	Lower limb movement & balance	III	Yes	80	-	26
	P7	Balance & right upper limb	I	No	2	4	17.6
	P8	Motor coordination & planning	-	Yes	44	-	12.3
	P10	Autism Spectrum Disorder	-	No	0.5	8	29
Occasional	P3	Balancing & jumping	II	No	3	4	27
	P9	Autism Spectrum Disorder	-	Yes	49	-	18.5
Unlikely	P1	General weakness & joint movement	V	Yes	-	-	53.25
	P2	Neurotypical development	-	No	2.6	6	42.8
	P4	Severe lower limb limitations	IV	Yes	-	-	54.2
	P5	Severe lower limb limitations	IV	Yes	-	-	25.3

We observed that children with a higher GMFCS level and, therefore, less mobility are less likely to be able to play the games using the dance mat and, as expected, require more active support from the therapists. Especially when they also suffer from cognitive disabilities. For the patients on the autism spectrum, the bigger challenges were having them interested in the games for long periods and the abstract perception of the connection between the button pressed on the mat and the movement of the car, but the therapist accompanying them believed these were all aspects that could improve in subsequent sessions as the children got more used to the games, and that the style of the games developed could be useful to train their focus.

For each game we collected data that was considered to be relevant to evaluate patient performance. For "Matchmat" (G1) and "Left or Right?" (G2), for each button that the child presses we save which button was pressed and the timestamp for that action, how much time the button was held, if it was the correct button, and we also calculate the Manhattan Distance (MD) of the button pressed to the target button. For the purpose of this paper we focused on the last stat as our success measure, the shorter the total Manhattan distance is, the better the child performed at the game. For the "Crazy Car" (G3) game, since the goal is to dodge obstacles for as long as possible the metric for success is the game play Time (s), the longer the better, but we also save each button pressed with an associated timestamp.

Because of the size of our sample, it is hard to make general assumptions about the collected data. Even so, taking another look at Table 2, we can see the average results each patient got in the games they played. In the column for Manhattan Distance (G1 MD#) of the "Matchmat", we can see that the values vary greatly from patient to patient and that the ones with more mobility constraints could not play it. Each of the abnormally high values are justified for a different reason for each patient. For P6 it was her slight cognitive deficit and even when she recognized the correct button, since she was playing while laying on a Pilates ball, she preferred to press the buttons closer to her because she didn't feel like stretching to reach buttons further away - this was one the aspects of her condition that therapist felt could be improved with the games. When his session started, Patient P8 was not very interested in playing the games and wanted to be held by the therapist leading to poor cooperation in the "Matchmat" game. Patient P9 was someone who had trouble processing new information and had trouble making the connection between the mat and computer screen. He was also much more interested in playing with the computer.

Only the four patients that managed to play "Matchmat" correctly were able to play "Left or Right?". We can see in the G2 MD# column that the Manhattan distance remained within understandable values, despite raising slightly. This was the result we expected since the interaction method is practically the same but the visual representation is a little more complex.

Finally, the last column of Table 2 (G3 Time (s)) present the time each patient managed to play the "Crazy Car" game without hitting an obstacle. The reason we some patients with less mobility achieving better results is because the patients with less mobility were playing the games, sitting down, with excessive help from the therapists accompanying them, which to a certain degree would nullify the benefits of having the child play the game. That is also one of the big reasons therapists said it is unlikely that they would be interested in integrating the games in the therapy sessions of these patients. Regardless of that, both the development and the expert team are optimistic with the obtained results and look forward to continue exploring this co-design process.

Analyzing patients P6, P7, P8, and P10, we observe the system was more suited for children that retained some mobility in at least one group of limbs. These children can harness more potential from the mat because they do not require a level of assistance that would essentially have therapists playing the games for them. It is obvious that children with typical cognitive development understand the games faster and better, but patient P6 shows that having a slight cognitive deficit is not an obstacle to interact with the games, as long as they can understand and recognize shapes, which is something they might even train while playing by having to recognize the symbols on screen repeatedly.

The different ways the children interacted with the dance mat also showed us it is a versatile game controller capable of adapting to the different needs of patients. The dance mat can act as a very flexible controller that can be used in a myriad of ways at a very affordable price, which would allow the parents to easily get one and help their children play the games at home. This expands the places where the system can be used, which can only benefit the end-user.

The participatory design based on a design thinking approach gave us valuable insights on the creation of exergames for children with disabilities using a dance mat, which can be summarized in the following main guidelines:

- Usability: The games can be used by children, who find the experience positive, and the therapists identify they might be mostly useful for children with motor and balance constraints. Therapists also consider the game to be able to potentially train focus in occupational therapy.
- Flexibility: The dance mat is a very robust and flexible controller, being important that the games can be highly configurable for each child's needs.
- Feedback: Whether positive or negative, it is one of the most important parts of the game. The patients must understand if they are performing the actions correctly and feel they are progressing.
- Tracing: Therapists want to register the progress and observe the history of the patient. It is also important to have an online platform to offload the burden of managing records from the game designers.

6 Conclusions and Future Work

In conclusion, throughout every stage, the therapists were very interested in this collaborative process. The participatory design allowed the developers to create a product that may prove to be useful to therapy in further studies. Going forward, patient selection must be done in a careful way to make sure children are not frustrated when playing the exergames. In general, every child seemed enthusiastic to play the games using the dance mat, but therapists believe the potential therapy benefits could only be harvested by children with a GMFCS level III, or lower, or with other children that retain some level of independence of movement in the upper and/or lower limbs.

The integration with an external framework [3] will be further explored to make use of the provided tools allowing real-time update of game settings and managing game result recordings. Finally, we must conduct a thorough evaluation with children from the clinic, but also with therapists and children from other clinics, who did not participate in the design.

References

1. Al-Nasri, I., Salim, S.: Using gamification to break barriers in physical therapy. Unive. West. Ont. Med. J. **87**(2), 9–11 (2019). https://doi.org/10.5206/uwomj.v87i2.1146
2. Aloba, A., et al.: Toward exploratory design with stakeholders for understanding exergame design. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–8. CHI EA 2020, ACM, New York (2020). https://doi.org/10.1145/3334480.3382784
3. Antunes, A., Madeira, R.N.: Play - model-based platform to support therapeutic serious games design. Procedia Comput. Sci. **198**, 211–218 (2022). https://doi.org/10.1016/j.procs.2021.12.230
4. Barendregt, W., Börjesson, P., Eriksson, E., Torgersson, O., Bekker, T., Skovbjerg, H.M.: Modelling the roles of designers and teaching staff when doing participatory design with children in special education. In: Proceedings of the 15th Participatory Design Conference. PDC 2018, ACM, New York (2018). https://doi.org/10.1145/3210586.3210589

5. Barrett, N., Swain, I., Gatzidis, C., Mecheraoui, C.: The use and effect of video game design theory in the creation of game-based systems for upper limb stroke rehabilitation. J. Rehabil. Assistive Technol. Eng. **3**, 2055668316643644 (2016). https://doi.org/10.1177/2055668316643644
6. Bonnechère, B., Omelina, L., Jansen, B., Rooze, M., Van Sint Jan, S.: Balance training using specially developed serious games for cerebral palsy children, a feasibility study. In: Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare, pp. 302–304. PervasiveHealth 2014, ICST, Brussels, BEL (2014). https://doi.org/10.4108/icst.pervasivehealth.2014.255332
7. Brooke, J.: SUS: A 'Quick and Dirty' Usability Scale. In: Usability Evaluation In Industry, pp. 207–212. CRC Press, Boca Raton (1996). https://doi.org/10.1201/9781498710411-35
8. Brown, T.: Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation. Harper Collins, New York (2009)
9. Burke, J.W., McNeill, M.D.J., Charles, D.K., Morrow, P.J., Crosbie, J.H., McDonough, S.M.: Optimising engagement for stroke rehabilitation using serious games. Vis. Comput. **25**(12), 1085–1099 (2009). https://doi.org/10.1007/s00371-009-0387-4
10. Bykbaev, V.R., Vélez, E.P., Guerra, P.I.: An educational approach to generate new tools for education support of children with disabilities. In: Proceedings of International Conference on e-Education Entertainment and e-Management, ICEEE 2011, pp. 80–83 (2011). https://doi.org/10.1109/ICeEEM.2011.6137847
11. Caro, K., Martínez-García, A.I., Kurniawan, S.: A performance comparison between exergames designed for individuals with autism spectrum disorder and commercially-available exergames. Multimedia Tools Appl. **79**(45), 33623–33655 (2020). https://doi.org/10.1007/s11042-019-08577-y
12. Caro, K., Morales-Villaverde, L.M., Gotfrid, T., Martinez-Garcia, A.I., Kurniawan, S.: Motivating adults with developmental disabilities to perform motor coordination exercises using exergames. In: Proceedings of 4th EAI International Conference on Smart Objects and Technologies for Social Good, pp. 183–189. Goodtechs 2018, ACM, New York (2018). https://doi.org/10.1145/3284869.3284914
13. Kelly, H., Howell, K., Glinert, E., Holding, L., Swain, C., Burrowbridge, A.: How to build serious games: Holding. Commun. ACM **50**, 44–49 (2007). https://doi.org/10.1145/1272516.1272538
14. Edmans, J., Gladman, J., Walker, M., Sunderl, A., Porter, A., Fraser, D.S.: Mixed reality environments in stroke rehabilitation: development as rehabilitation tools. Int. J. Disabil. Hum. Dev. **6**(1), 39–46 (2007). https://doi.org/10.1515/IJDHD.2007.6.1.39
15. Frauenberger, C., Good, J., Alcorn, A.: Challenges, opportunities and future perspectives in including children with disabilities in the design of interactive technology. In: Proceedings of the 11th International Conference on Interaction Design and Children, pp. 367–370. IDC 2012, ACM, New York (2012). https://doi.org/10.1145/2307096.2307171
16. Hernandez, H.A., Ye, Z., Graham, T.N., Fehlings, D., Switzer, L.: Designing action-based exergames for children with cerebral palsy. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1261–1270. CHI 2013, ACM, New York (2013). https://doi.org/10.1145/2470654.2466164
17. Houtrow, A., et al.: Prescribing physical, occupational, and speech therapy services for children with disabilities. Pediatrics **143**(4), e20190285 (2019). https://doi.org/10.1542/peds.2019-0285

18. Howard, T., et al.: Designing an app to help individuals with intellectual and developmental disabilities to recognize abuse. In: The 23rd International ACM SIGACCESS Conference on Computers and Accessibility. ASSETS 2021, ACM, New York (2021). https://doi.org/10.1145/3441852.3471217
19. Hwang, S., et al.: How game balancing affects play: player adaptation in an exergame for children with cerebral palsy. In: Proceedings of the 2017 Conference on Designing Interactive Systems, pp. 699–710. DIS 2017, Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3064663.3064664
20. Jacob, J., Lopes, A., Nóbrega, R., Rodrigues, R., Coelho, A.: Towards player adaptivity in mobile exergames. In: Cheok, A.D., Inami, M., Romão, T. (eds.) ACE 2017. LNCS, vol. 10714, pp. 278–292. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76270-8_20
21. Kozyavkin, V.I., Kachmar, B.O., Terletskyy, O.I., Kachmar, O.O., Ablikova, I.V.: Stepping games with dance mat for motor rehabilitation. In: 2013 International Conference on Virtual Rehabilitation, ICVR 2013, pp. 174–175 (2013). https://doi.org/10.1109/ICVR.2013.6662108
22. Lo, J.L., Chi, P.Y., Chu, H.H., Wang, H.Y., Chou, S.C.T.: Pervasive computing in play-based occupational therapy for children. IEEE Pervasive Comput. **8**(3), 66–73 (2009). https://doi.org/10.1109/MPRV.2009.52
23. Madeira, R.N., Antunes, A., Postolache, O., Correia, N.: Serious...ly! just kidding in personalised therapy through natural interactions with games. In: Cheok, A.D., Inami, M., Romão, T. (eds.) ACE 2017. LNCS, vol. 10714, pp. 726–745. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76270-8_50
24. Palisano, R., Rosenbaum, P., Walter, S., Russell, D., Wood, E., Galuppi, B.: Development and reliability of a system to classify gross motor function in children with cerebral palsy. Dev. Med. Child Neurol. **39**(4), 214–223 (1997). https://doi.org/10.1111/j.1469-8749.1997.tb07414.x
25. Patel, D.R.: Therapeutic interventions in cerebral palsy. Indian J. Pediatr. **72**(11), 979–983 (2005). https://doi.org/10.1007/BF02731676
26. Putnam, C., Puthenmadom, M., Cuerdo, M.A., Wang, W., Paul, N.: Adaptation of the system usability scale for user testing with children. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–7. CHI EA 2020, ACM, New York (2020). https://doi.org/10.1145/3334480.3382840
27. Sipiyaruk, K., Gallagher, J.E., Hatzipanagos, S., Reynolds, P.A.: A rapid review of serious games: from healthcare education to dental education. Eur. J. Dent. Educ. **22**(4), 243–257 (2018). https://doi.org/10.1111/eje.12338
28. Thompson, D., et al.: Serious video games for health: how behavioral science guided the development of a serious video game. Simul. Gaming **41**(4), 587–606 (2010). https://doi.org/10.1177/1046878108328087
29. van Gelder, L., Booth, A.T., van de Port, I., Buizer, A.I., Harlaar, J., van der Krogt, M.M.: Real-time feedback to improve gait in children with cerebral palsy. Gait Posture **52**, 76–82 (2017). https://doi.org/10.1016/j.gaitpost.2016.11.021
30. Zablotsky, B., et al.: Prevalence and Trends of Developmental Disabilities among Children in the United States: 2009–2017. Pediatrics **144**(4), e20190811 (10 2019). https://doi.org/10.1542/peds.2019-0811

Designing Hearing Aids to Mitigate Perceived Stigma Associated with Hearing Impairment

Guanhua Sun, Leila Aflatoony(✉), and Wendell Wilson

School of Industrial Design, Georgia Institute of Technology, Atlanta, USA
gsun63@gatech.edu, leila.aflatoony@design.gatech.edu

Abstract. Assistive technologies support people with disabilities in living independently. However, perceived stigma towards using assistive technologies can lead to issues such as non-acceptance or abandonment of these technologies. This study deconstructs stigma perception in people with hearing impairment supported by social psychology and design elements. Users' feelings towards hearing aid products and their stigmatization are illustrated through related literature and findings from an empirical study. The study revealed the stigmatization associated with hearing impairment and hearing aids, and strategies to overcome related barriers. We then developed a series of hearing aid prototypes in a design workshop and evaluated their efficacy in mitigating stigma in people with hearing impairments. We discuss our findings on stigma threats in hearing aid products, as they relate to the sources of stigma, strategies to eliminate stigma. We suggest considering inclusive design principles to develop mainstream hearing aid products.

Keywords: Human-Centered Design · Stigma · Hearing Aids · Hearing Impairment · Assistive Technology

1 Introduction and Related Work

Peoples' ability to hear is one of the most significant tools to explore the world since it provides them with vital information about their surroundings. Hearing loss affects more than just the ability to hear and has become more frequent in recent years across all age categories. As a result of hearing loss, people may feel socially isolated and communicate ineffectively, even be stigmatized, with their mental health and quality of life detrimentally influenced [14]. Assistive technology (AT) is any item, piece of equipment, software program, or product system that is used to help people with disabilities increase, maintain, or improve their functional abilities across a variety of living domains [2, 17]. Despite the abundant advantages of AT devices, people with disabilities may refuse to utilize these products due to various reasons, including cost, discomfort, a sense of foreignness, stigmatization, social rejection, and embarrassment [7]. The use of assistive technologies can be affected by users' acceptance and acceptability of the products [18].

For a person living with disability, stigmatization is often a reality having varying effects, including, but not limited to, 1) less treatment; 2) disrupted social relations;

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 528–539, 2023.
https://doi.org/10.1007/978-3-031-34586-9_35

3) personal avoidance, anxiety, and depression; and 4) disordered self-image and poor self-esteem [6]. People of any age can feel stigmatized by AT devices that represent their loss of physical functions. The majority of studies of ATs are primarily concerned with functionality and usability, ignoring the value of self-expression and social context in determining the long-term adoption of these technologies [7]. Social context is an essential element that influences people's attitudes towards those with specific impairments and use of AT devices. Many psychosocial factors, including personality, response to disability, and the environment or social milieu where the technology is applied, affect the acceptability of AT. Social acceptability has been identified as one of the critical elements impacting whether a person uses a particular AT device [9]. People's acceptance of AT products is an important factor contributing to higher volume manufacturing and engaged use, thus reducing the risk of product abandonment [11].

Currently, design has moved to human-orientation shifting from designing for users to designing with users. Considerable emphasis has been placed on the exploration of the relationship between AT products and stigma, and several design strategies are used to achieve destigmatization [15]. The first strategy is to reshape the societal context. As a response to product-related stigma in this context, one should choose interventions that either produce fundamental changes in attitudes and beliefs or change power relations that underlie the ability of domain groups to act upon their attitudes and beliefs [5]. The second strategy is to reshape the meaning of the products. Factors such as shapes, material qualities and other sensory demonstrations, are considered as design elements. Through their presence coupled with other sensory elements, a product has the potential of imposing a stigma on its users or wearers, both physically and psychologically [10]. The means of managing stigma can be divided into three categories: disguising the stigmatizing features, shifting the attention from the stigmatizing features to other features, and transforming stigmatizing elements into features that convey prestige or a higher social status [4]. To identify factors that affect the perception of stigma, this study conducted an online survey and semi-structured interviews to ascertain whether people with hearing loss faced stigmatization and identify the origin of stigmatization. We then developed proof-of-concept porotypes of hearing aids to evaluate their benefit in potential reduction of stigmatization in people with hearing impairment.

2 Methodology

2.1 Online Survey

A ten-question survey was dispersed to Facebook groups, including Hearing Aid Forum, Community for the Deaf and Hard of Hearing, and Living with Hearing Loss Group. In the first phase, 98 people took the survey. Basic demographic information, including age, gender, and the cause of hearing impairment, was collected to understand the difference between various user groups. Participants shared their feelings, current habits, and concerns regarding hearing aid through open-ended survey questions. We coded their responses (PS-1 to PS-98) to protect their identities. This study has been reviewed and approved by the Institutional Review Board (IRB) at Georgia Tech.

2.2 Semi-structured Interview

Remote interviews were conducted after the completion of the survey to collect more in-depth data concerning participants' perception of stigmatization. Seven participants were recruited through Facebook groups, who had completed the online survey and were willing to share more about their experiences with hearing loss and hearing aids. The semi-structured interview centered around users' feelings regarding hearing loss, their experiences with current hearing aids, their perception with stigmatization when wearing assistive products, and their expectations towards next generation products. Participants included 5 females and 2 males ranging in age from 35–70. All interviewees used hearing aids in their daily lives. Their names were coded (PI-1 to PI-7) to protect their identities. Answers from the survey and interviews were organized into an affinity diagram to identify emerging themes and underlying codes.

3 Results

3.1 Stigmatization Associate with Hearing Impairments

Communication. Communication, and self-perception [16] in people's daily lives are subject to the influence of hearing loss, thus resulting in the occurrence of stigmatization. Analysis of the survey results revealed that almost every participant had difficulties in communication, even those with hearing aids. Hearing aids improved their hearing by amplifying sounds around them, but at the same time, noises were amplified as well. It was revealed in the online survey that, given the noise from hearing aids, approximately 37% of the participants refused to wear any hearing aid. Additionally, PS-11 said that hearing aids did not reduce noises, and her voice resonated through traditional hearing aids: "*I tried a traditional hearing aid but did not like it because the sound was muffled like it was underwater, but it was also way too loud regardless of how much I turned it down. Plus, my voice resonated through the hearing aids*".

Although newer technologies have effectively reduced noise, some people suffering from hearing loss still have trouble communicating. They cannot follow every word when talking, especially when the pitch is too low (males) and too high (children). In the survey, some participants complained that communication required more effort on their part, which made them less willing to talk with others. Daily communication is still a difficult task, and barriers in communication can bring negative emotions to people with hearing impairment. As PI-6 stated in the interview,

> "*I worked in the library, so I need to talk with different people. I feel embarrassed if I cannot hear them clearly or cannot understand what they say. There are lots of men in my previous work, and their pitch is very low, so I cannot hear them. I need to talk to different people. And if I couldn't hear them, they got frustrated, and I got frustrated. I will show less confidence about that*".

To tackle these issues, participants shared strategies they use such as speech-text software, sign language, and lip-synch. However, PS-70 noted in the survey that "*with the use of face masks I now realize I was relying more on lip-reading than I thought*".

People's hearing loss significantly affects their ability to participate in social activities, which derives from their inability to converse with more than one person at a time as PS-57 shared: "*When I talk to a group of people, I cannot hear them clearly, and I need to interrupt them. It is very awkward and impolite*". Lack of ability to communicate in group activities cause sensation of isolation as one of the stigmatizing factors. For example, PI-1 mentioned in the interview: "*In some social activities, everything is emphasized, including hearing loss and my isolation from crowds. Everyone involved in this activity would know that I have a hearing impairment*".

Self-perception. Functional disability affects people's self-perception, which can be referred to as stigmatization. In the survey, participants under the age of 45 were more concerned about their hearing loss, which hampered their personal development, and they felt stigmatized. People with hearing disabilities have long refused to mention or try to talk about their disabilities, and they are sensitive to words not directly related to their bodies, such as "deaf," since they feel that these words carry malice. Participants shared ideas about how their perceptions were affected by hearing loss and hearing aids, and how they thought of others' perceptions of them. For example, PS-45 stated that:

> "*Some people think that deaf people are not smart, so they treat me differently. I mean with my training and the help of hearing aids; the hearing function can be restored a lot. I don't need their special treatment. These treatments just reminded me that I was different from others, and I would feel stigmatized*". She also noted that she felt hearing loss "*diminishing one's authority*" and that authority mattered considerably in her work role.
>
> "*I am a teacher, but sometimes it is hard for me to follow students. I cannot have eye contact with students. I need to look at their lips. That's why there was a certain distance between the students and me. Parents may not trust me, a teacher with hearing loss. They would doubt my capacity to communicate effectively with pupils and the quality of my instruction*".

Overall, hearing loss itself can cause stigma, mostly in the form of obstacle to communication and altered self-perception. The survey and interview results revealed that, adults can expose to more psychological strain while dealing with problems brought by hearing loss. When people had been accustomed to a certain way of life, a sudden shift in their sensory system could be jarring. This sensation of psychological discord was accompanied by challenges in daily life and manifested as the perception of stigma. As PS-4 shared, children were more receptive to hearing loss and hearing aid devices: "*for children/babies, they underwent surgery at a very young age, so they will not suffer hearing problems during growth. And they could accept hearing devices more easily than teenagers and adults*".

3.2 Stigmatization Associated with Hearing Aids

The primary reason for perceived stigmatization shared by participants with hearing loss was associated with wearing hearing aid products. From the survey, about 87% of participants stated that hearing aids could help them improve their hearing in most situations, but some still refused or struggled with hearing aids. While hearing aids amplify the sounds around people, they also amplify/accentuate one's personal defects, i.e., a person with hearing impairment will be more easily noticed so several participants shared they do not want to expose their impairment to the public. PI-3 stated that:

> "*Without my hearing aids, I wouldn't be able to communicate. They work well. I would compare this with visual impairment. If you have a visual impairment, you can wear glasses. And you still look like a normal person. But if you wear a hearing aid product, everyone will know that I have a problem with hearing*".

She also remarked that she, like many others, tended to conceal her hearing aids. The invisible design conceals not only the hearing aids but also the stigmatization that surrounds them:

> "*Anyone who says they are not embarrassed wearing hearing aids is lying. I choose the CIC variant as nearly invisible. In my case, it makes me an 'old man' before my time. It might be better for women with longer hair. They can hide BTE aids better.*"

Participants shared their concerns about products in the actual use process, not only for the practical support, but also for the humanistic care behind additional items. People prefer hearing aid items that hide their impairments and make effort to hide the items. For example, PI-7 shared how she helped her son hide his hearing aids:

> "*My son had a cochlear implant when he was very young. I was worried that he would receive discrimination and feel inferior because of his hearing aids, and I would help him hide them in his hair. He now has a haircut that looks like a little hat to block his hearing aids*". She stated she made many compromises to hide the hearing aid: "*Since his cochlear implant surgery, my kid has had the same hairstyle, and it hasn't changed in years. Even though it's hot, many little girls may leave their hair long for a long time to conceal their cochlear implants.*"

3.3 Strategies to Overcome Stigma

Not all participants were impacted by the perception of stigma. For example, some participants decorated their hearing aids instead of hiding them and gave their hearing aids a different definition, which was easier for others to accept their functional impairments. As shared by PI-7, her son designed many stickers for his hearing aids. He would show his hearing aid to his friends and tell the children around him that it was a gift from an anime character:

> "*He figured out how to make his hearing aids look cool. It was hard for him to make friends because of his hearing before. But ever since he got into watching anime, he started to think about how he could make himself look like anime characters. So, he started designing anime peripherals and combining these with his hearing aids. He showed his design to the kids around him and soon attracted many of them. This was also an opportunity for him to integrate into the circle of his peers slowly*".

PI-6 also stated that a considerable number of people had grown more accepting of ear electronics due to the presence of a wide variety of Bluetooth headsets. One of the participants eagerly expressed her expectation of the next generation of hearing aid to more closely resemble electronic products, to be more beautiful while maintaining functionality, thus allowing the technology to be blended into fashion trends in modern society: "*Although I can hide my hearing aid under my hair, if it is like the AirPods, I just look like a normal person*".

4 Iterative Prototyping of Hearing Aids

4.1 Design Workshops

A design workshop was adopted in the project to develop design alternatives and elements that effectively reduce stigmatization in potential users. The research team created a design workshop with three design students. A design workshop was chosen to assist designers better understanding the usage scenarios encountered by hearing loss people. Before the establishment of the workshop, existing hearing aid models on the market were compared and their benefits and drawbacks were correspondingly evaluated. There are many different types of hearing aids available on the market, and their suitability is determined by the degree of hearing loss (see Fig. 1).

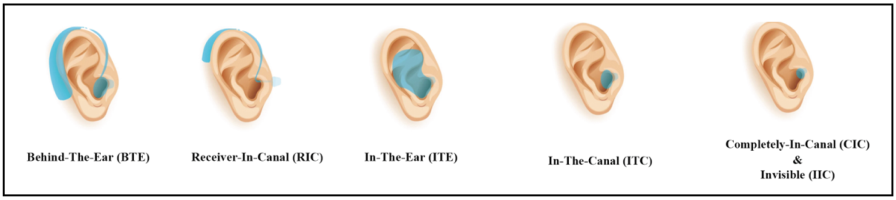


Fig. 1. Different types of hearing aids. Images are taken from hkincus.com.

The workshop was comprised of several activities, including the production of storyboards, 3D prototypes, and debriefing interviews. Each participant was asked to complete at least one storyboard, in which a usage scenario was predefined. The target users were further subdivided based on the defined scenario. Participants illustrated the action in the storyboard and considered the functional design and product form of a hearing aids. Participants were asked to draw sketches and create 3D models based on their proposed scenarios. They were provided with various modeling materials such as wire, pipe cleaners, and clay. Designers created different design scenarios and use cases. For example, a designer mainly focused on female users and combined hearing aid products with fashionable designs based on BTE Model. Other designers explored hearing aids to encourage participation in social activities and broaden the social circle of those with hearing impairments, thereby lowering stigma (see Fig. 2).

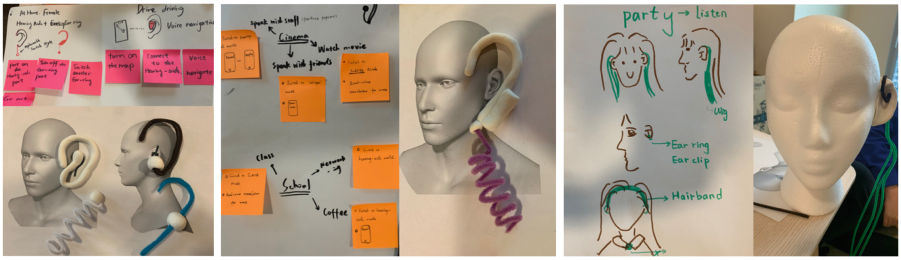


Fig. 2. Storyboard and prototype developed by designers in the design workshops.

4.2 Prototyping

Based on the result from the workshop and the preliminary empirical findings from the survey/interview, we developed two design prototypes. The first strategy was to reshaping product by de-identification. This strategy can be divided to into camouflage and diversion of attention categories. Designers would utilize translucent or skin-colored materials in the camouflage method. The device was concealed consumer electronics so the user would not draw attention while wearing the product. There were two design directions available in the first strategy: the first was a binaural independent Bluetooth earphone based on the ITE Model, while the second was a neckband headphone based on the RIC Model (see Fig. 3). The second strategy was to reshaping product by identification. Personalization allows users to choose or change the product in such a way that it compliments and expresses their identity, for example, the adding of lifestyle components. Compared to the design of traditional hearing aids, this product was defined as smart jewelry to transform them into fashionable wearables (see Fig. 3).

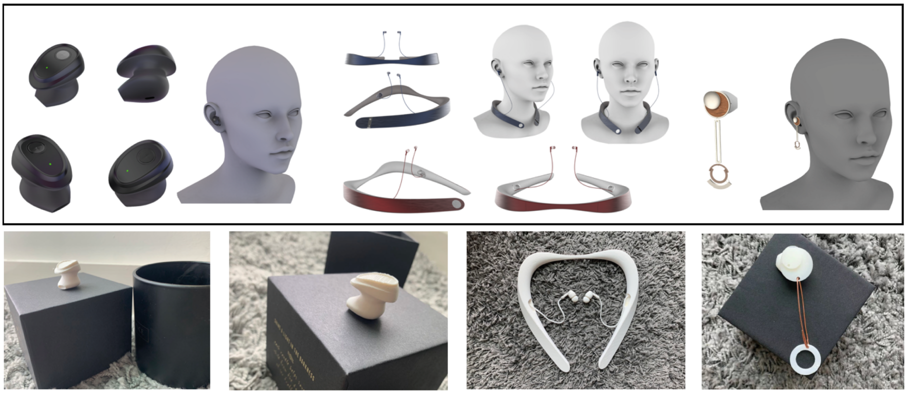
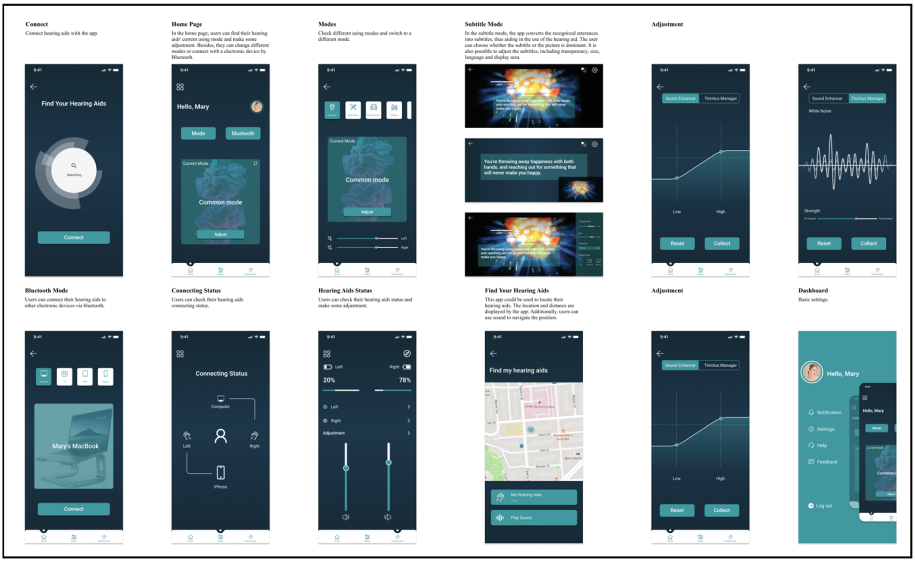


Fig. 3. From left to right: Binaural independent Bluetooth earphone based on the ITE Model; Neckband headphone based on the RIC Model; and Smart jewelry based on the ITE Model.

4.3 Interface Design

The interface design (see Fig. 4) is used to configure hearing aids. Users need to connect their hearing aids with the app via Bluetooth when they log in. They are allowed to check current hearing aids moods and make adjustments. They can also switch between different modes and link their aid with other electronic devices such as cell phones, iPads, and laptop computers. A sound enhancer and tinnitus management tool are used to fine-tune the sound output and make the sound output more acceptable for the hearing impaired. Based on the literature review and design workshop, different scenarios are defined in the interface design to satisfy the demands of various circumstances, including, modes for noisy places, transportation as well as speech to text option and microphone mode focuses on more private communication. The multi-functional hearing aid design avoids the need for users to switch between products while using them. Users can also monitor the current connection status and usage of their hearing aids, making it simple to manage their hearing aids.



Connect
Connect hearing aids with the app.

Home Page
In the home page, users can find their hearing aids' current using mode and make some adjustment. Besides, they can change different modes or connect with a electronic device by Bluetooth.

Modes
Check different using modes and switch to a different mode.

Subtitle Mode
In the subtitle mode, the app converts the recognized utterances into subtitles, thus aiding in the use of the hearing aid. The user can choose whether the subtitle or the picture is dominant. It is also possible to adjust the subtitles, including transparency, size, language and display area.

Adjustment

Bluetooth Mode
Users can connect their hearing aids to other electronic devices via bluetooth.

Connecting Status
Users can check their hearing aids connecting status.

Hearing Aids Status
Users can check their hearing aids status and make some adjustment.

Find Your Hearing Aids
This app could be used to locate their hearing aids. The location and distance are displayed by the app. Additionally, users can use sound to navigate the position.

Adjustment

Dashboard
Basic settings.

Fig. 4. Interface design associated with hearing aids.

5 User Study

A user study was conducted after the prototypes were fabricated. The goal of the user study was to evaluate whether the proof-of-concept prototypes could reduce stigmatization and whether the mental application model established by the users matched conceptual design of the porotypes. Five users who had previous experiences with hearing loss and hearing aids were recruited, four of whom were female, and one was male.

In the user study, the participants were assigned to engage with both the mobile application and the 3D-printed prototypes. To guide their interactions, participants were asked to complete 7 tasks that had been developed for the user study:

- Participants wore different prototypes one by one.
- Participants adjusted the prototype to fit their ears.
- Participants connected the hearing aids with the App.
- Participants adjusted the volume of hearing aids using the interface prototype.
- Participants changed different usage modes through the interface prototype.
- Participants checked hearing aids connection status.
- Participants use the APP to find their hearing aids.

We observed their tasks and encouraged them to think out loud and describe their experience as they were exploring the prototypes. Following the user study, the participants completed a questionnaire consisting of Likert scale questions to evaluate their satisfaction with the design prototypes. After completing the questionnaire, participants were asked to voice their sentiments toward the system and evaluate the usefulness of the prototype and discuss how the conceptual model of the system deviated from their own. We asked questions such as, what do you think of our concept? Are you willing to wear them in your daily life? Are you willing to recommend these prototypes to your friends?

5.1 Results

All participants completed a post-usability survey to evaluate their overall satisfaction, which included seven questions, rated on a scale of 1 = strongly disagree to 5 = strongly agree (Fig. 5). We calculated the mean for each answer provided and found the overall average number for participants satisfaction rate (Q1 to Q7) is 4.085 out of 5.

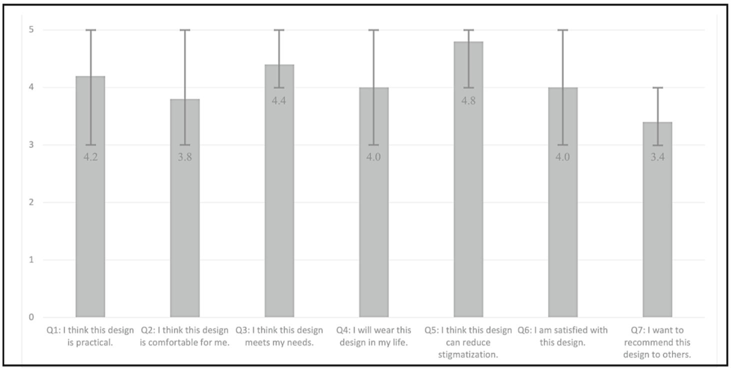


Fig. 5. The post-survey results on three prototypes' satisfaction rates show the calculated average mean per question. Error bars represent the maximum and minimum values.

All five participants expressed that the prototypes could reduce the perception of stigma. They believed that disguising stigmatizing features was the most effective approach. People were less concerned about whether assistive technologies were worn in the ear when they were hidden behind common consumer products. As P2 stated: "*I would think this is the best design solution! People wear headphones, maybe for listening to songs, maybe for navigation. If I wear the same product, people will think I'm listening to a song and won't notice if I'm using a hearing aid*". They also thought the smart jewelry design is appealing and regarded it as a demonstration of personality. These design concepts still have some drawbacks. P1 claimed that the friction of her hair may affect the effectiveness of the hearing aid. She had to tie her hair at the back of her head while she wore her hearing aids. P4 expressed his concern about the wearability of smart jewelry solutions. The hearing aid might fall out in case of quick movements: "*I don't think it's the accuracy of 3D printing. I mean, it wasn't able to fit my ears perfectly. Will he fall out when I'm working out? Of course, a lot of headphones are not designed to ensure that they do not fall out completely during exercise*".

The jewelry requirements for women varied greatly in different situations, they hoped that basic kits could be available for users to design their hearing aids. They hoped the design could give users more color schemes to meet the preferences of different people. Color is a design element that can reveal the user's personality. "*Women's requirements for accessories will change depending on the occasion. I don't think a single smart jewelry design can fully satisfy my social needs*". P5 expressed a desire for the jewelry's appearance to be more exaggerated. As a social opportunity point, the current design is a bit conservative: "*This design may be a bit ordinary for being a talking point. It may attract attention, but far less than a more exaggerated design*".

Additionally, Participants shared that the numerous modes created in the application, could be a useful for people wearing hearing aids and help them better participate in social events. Bluetooth mode could make it easier for them to utilize a range of electronic devices in a variety of scenarios without being overwhelmed. Hearing aids could become more powerful due to their systematic design solutions. The design prototypes could be integrated into people's daily lives life, breaking the constraints of the original AT products. These design concepts could offer them substantial assistance in a variety of settings and make them more confident participating in social activities: "*The design of the usage modes is really helpful! I can participate in different social scenarios, encouraging me to be proactive socially*". Furthermore, the subtitle model was found to be useful. Even with hearing aids, many people with hearing loss struggle to communicate with others, and some still rely on visual signals to read information. Subtitle mode could be customized to fit several situations. As P5 stated: "*The subtitle mode allows me to study and work more efficiently. The subtitle mode is designed with details that consider the difference between users and is very useful*".

6 Discussion

The perception of hearing aids as being stigmatized is a key impediment to hearing loss people's development. Much of the stigma among those with hearing loss is related to wearing hearing aids, rather than the hearing loss itself [19], defined as the "hearing aids effect" first mentioned in 1977 by Blood, Blood, and Danhauer [1]. This study revealed stigma problems from different perspectives, including social behavior and self-esteem. Social participation is defined as a person's involvement in social activities that provide them with interaction with others in the community [6]. Social engagement, as interactions with potential ties in real life, provides individuals with a coherent and consistent sense of role identity, companionship, and sociability as well [3]. People with hearing impairments have a substantially lower level of social participation due to their functional deficiencies and designing appropriate hearing aids can improve their social involvement and activities. Additionally, self-perception, or people's personal views about themselves, has a significant impact on the type of activity people engage in, the efforts they will put into that activity, and the likelihood that they will engage in that activity [12]. Self-perception, along with the judgment of individuals from others, helps assess the quality of social relationships [13].

Designing hearing aids should consider improving self-perception and promoting social engagement in people with hearing impairment. By deconstructing the using process and refining the system of hearing aids, the perception of stigma can be reduced regarding design elements and functionality. Designers should consider adopting an empathic and human-centered design approach to designing hearing aids (and other ATs) to better understand people living with impairments needs and enhance their well-being by proposing mainstream product solutions that improve their self-esteem and social interaction. The social acceptability of using hearing aids can be affected by their functional flaws so designers should be aware that these products should psychologically assist people with disabilities.

Adults with hearing loss are more likely to have a poorer personal and family socioeconomic position, leaving them more frequently exposed to negative life events, and unhealthy behaviors, thus resulting in increased physical and psychological stress. They may experience rage and irritation on a regular basis, in adjusting to a world that is not built for them. We suggest designing hearing aids following inclusive design principles, which emphasizes the notion of mainstream products, which eliminate the need for ATs in the definition: "*The design of mainstream products and/or services that are accessible to, and usable by, as many people as reasonably possible on a global basis, in a wide variety of situations and to the greatest extent possible without the need for special adaptation or specialized design*" [8]. Considering product semantics can be useful in solving stigma-related design concerns. This study has some limitations. Firstly, during the COVID time, part of the design research and usability tests were conducted remotely, so important information might have been overlooked. Second, given that the participants were mostly women, the final design outcome may not be fully representative of the entire community. Future research needs to be conducted to further develop and explore mainstream ATs and hearing aid design alternatives to reduce or eliminate perceived stigma in people with hearing impairments.

Acknowledgments. We would like to express our gratitude to all the hearing loss participants and their families who generously contributed their time and efforts to this study and offered valuable information for this work.

References

1. Blood, G.W., Blood, I.M., Danhauer, J.L.: Listeners' impressions of normal-hearing and hearing-impaired children (1978)

2. Cook, A.M., Polgar, J.M.: Assistive Technologies-e-Book: Principles and Practice. Elsevier Health Sciences (2014)
3. Gao, J., Hu, H., Yao, L.: The role of social engagement in the association of self-reported hearing loss and health-related quality of life. BMC Geriatr. **20**(1) (2020). https://doi.org/10.1186/s12877-020-01581-0
4. Jacobson, S.: Overcoming the stigma associated with assistive devices. In: Proceedings of the 7th International Conference on Design and Emotion, no. Figure 1, pp. 3–4 (2010)
5. Krefting, L.: The culture concept in the everyday practice of occupational and physical therapy. Phys. Occup. Ther. Pediatr. **11**(4), 1–16 (1991). https://doi.org/10.1080/J006v11n04_01
6. Levasseur, M., Richard, L., Gauvin, L., Raymond, É.: Inventory and analysis of definitions of social participation found in the aging literature: proposed taxonomy of social activities. Soc. Sci. Med. **71**(12), 2141–2149 (2010). https://doi.org/10.1016/j.socscimed.2010.09.041
7. Marti, P., Recupero, A.: Is deafness a disability?: Designing hearing AIDS beyond functionality. In: C and C 2019 - Proceedings of the 2019 Creativity and Cognition, pp. 133–143 (2019). https://doi.org/10.1145/3325480.3325491
8. Normie, L.R.: BS 7000-6:2005 design management systems – part 6: managing inclusive design by British standards. Gerontechnology **4**(3) (2005). https://doi.org/10.4017/gt.2005.04.03.012.00
9. Pippin, K., Femie, G.R.: Designing devices that are acceptable to the frail elderly: a new understanding based upon how older people perceive a walker. Technol. Disabil. **7**, 93–102 (1997). https://doi.org/10.3233/TAD-1997-71-211
10. Pullin, G., Higginbotham, J.: Design meets disability. Augment. Altern. Commun. **26**(4), 226–229 (2010). https://doi.org/10.3109/07434618.2010.532926
11. dos Santos, A.D.P., Ferrari, A.L.M., Medola, F.O., Sandnes, F.E.: Aesthetics and the perceived stigma of assistive technology for visual impairment. Disabil. Rehabil. Assist. Technol. (2020). https://doi.org/10.1080/17483107.2020.1768308
12. Shapka, J.D., Khan, S.: Self-perception. In: Levesque, R.J.R. (ed.) Encyclopedia of Adolescence, pp. 3406–3418. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-33228-4
13. Stinson, M.S., Whitmire, K., Kluwin, T.N.: Self-perceptions of social relationships in hearing-impaired adolescents (1996)
14. Tambs, K.: Moderate effects of hearing loss on mental health and subjective well-being: Results from the Nord-Trøndelag hearing loss study. Psychosom. Med. **66**(5), 776–782 (2004). https://doi.org/10.1097/01.psy.0000133328.03596.fb
15. Vaes, K., Stappers, P.J., Standaert, A., Desager, K.: Contending stigma in product design using insights from social psychology as a stepping stone for design strategies. In: 8th International Conference on Design and Emotion: Out of Control – Proceedings (2012). https://doi.org/10.13140/2.1.3738.8800
16. Wallhagen, M.I.: The stigma of hearing loss. Gerontologist **50**(1), 66–75 (2010). https://doi.org/10.1093/geront/gnp107
17. What is AT? Assistive Technology Industry Association. https://www.atia.org/home/at-resources/what-is-at/. Accessed 31 July 2022
18. Wilkinson, C.R., Angeli, A.: Applying user centred and participatory design approaches to commercial product development. Des. Stud. **35**(6), 614–631 (2014). https://doi.org/10.1016/j.destud.2014.06.001
19. Zaitzew, C.M.: IdeaExchange@UAkron understanding the stigma of hearing loss and how if affects the patient and treatment process recommended citation http://ideaexchange.uakron.edu/honors_research_projects/402

Demands on User Interfaces for People with Intellectual Disabilities, Their Requirements, and Adjustments

Melinda C. Braun[1,2](✉) and Matthias Wölfel[1,2]

[1] Karlsruhe University of Applied Sciences, 76133 Karlsruhe, Germany
{melinda.braun,matthias.woelfel}@h-ka.de
[2] University of Hohenheim, 70599 Stuttgart, Germany

Abstract. Information and communication technologies are ubiquitous in today's society. They have the potential to enhance the life of its users in various areas, especially the life of people with intellectual disabilities (ID). Unfortunately, natural user interfaces are often too complicated to use and not adapted to the varying needs of every user group. A possible improvement can be achieved by adapting the respective user interface to the abilities and skills of the respective user(s). Therefore, this study evaluates currently available interface types and their adaptation possibilities and requirements for people with ID. 116 individual solutions and prototypes were tested with 41 participants. We found that interfaces with pointing gestures are currently the preferred interface type for most people with ID, as this input type is used in most technologies today and provides the most accessibility features and possible adaptations. Other input types, such as voice or object interaction, offer great potential for people with disabilities and ID, but are currently more difficult to adapt to the individual needs of users with ID.

Keywords: natural user interfaces · consumer technologies · interface adaption · accessibility · intellectual disabilities

1 Introduction

Information and communication technologies (ICT) are ubiquitous in today's society and have become an essential part of people's daily lives [1]. They have the potential to enhance the life of its users in various areas, especially the lives of people with intellectual disabilities (ID) [2]. Unfortunately, natural user interfaces are often too complicated to use and not adapted to the varying needs of every user group. While design guidelines like "universal design" or the EU Directive 2019/882 (on accessibility requirements for products and services) [3] already exist, which is likely to continue to improve the accessibility of ICT, people with ID have very individual limitations and abilities. This makes it

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 540–556, 2023.
https://doi.org/10.1007/978-3-031-34586-9_36

difficult for developers or manufacturers of digital technologies to include every unique physical and intellectual difference.

An analysis of the current accessibility status of the natural user interface types "touch, voice, and touchless" showed a lot of existing problems when used by people with ID, but also a great potential for improvement. People with ID currently have difficulties in "accessing, selecting, or using different types of interfaces" [4]. This is the reason for the so-called digital divide that has formed between people with ID and the "regular" user [2]. Because of the underlying potential of ICTs, it is needed to reduce the digital divide and increase participation for users with special needs. A possible solution or improvement of this problem can be achieved by adapting the respective user interface to the abilities and skills of the respective user(s). Adaptations and customizations of technologies and interfaces can have different levels of complexity to implement, for this reason we created a continuum that shows the level of possible adaptation to modern mainstream technologies and assistive technologies.

The focus of this research is primarily on cost-effective and easy-to-use information and communication technologies, e.g., smartphones, tablets or smart home devices, which we refer to as consumer technologies in the following. The definition by [5] describes assistive technology devices as "any item, piece of equipment or product system whether acquired commercially off the shelf, modified, or customized that is used to increase, maintain or improve functional capabilities of individuals with disabilities", there is no clear distinction between assistive or consumer devices. For this reason, we distinguish consumer technology devices from assistive technology devices by the fact that they are not primarily and exclusively developed for people with disabilities, but for the general public. Consumer and assistive technologies can therefore be divided into different categories, depending on the degree of focus on people with disabilities. We created a continuum in which consumer technologies and assistive technologies represent the two sides (left and right) of the spectrum (see Fig. 1). In the middle, there are mixed forms between the two types of technologies, for instance with different degrees of adaptation. The categories were then coded and named as follows:

- **1. Consumer technologies:** regular consumer hardware, without special assistive software or hardware adaptations, e.g., an online banking-app used on a regular tablet or smartphone.
- **1A. Consumer technologies with operating system features that can improve accessibility:** regular consumer hardware that is more accessible through operating system settings, e.g., usability aids, larger fonts.
- **1B. Consumer technologies running assistive software:** consumer hardware combined with additional assistive software, e.g., an app that was developed for people with disabilities used on a tablet or smartphone.
- **1C. Consumer technologies with adaptations:** consumer hardware combined with hardware adaptions, e.g., adapting and customizing a smartphone interface with different sensors or additional buttons.
- **2. Special assistive devices:** special assistive hardware, developed for people with disabilities e.g., speech generating devices, special interfaces.

Consumer hardware is primarily marketed and distributed via the mass market while assistive hardware is marketed and distributed via health care providers.

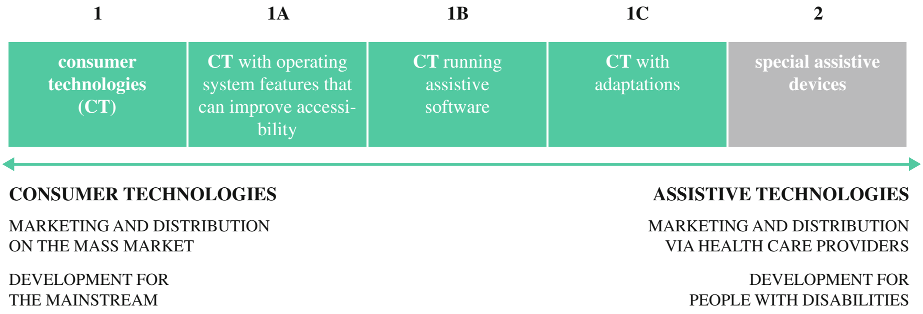

1
1A
1B
1C
2
consumer technologies (CT)
CT with operating system features that can improve accessibility
CT running assistive software
CT with adaptations
special assistive devices
CONSUMER TECHNOLOGIES
MARKETING AND DISTRIBUTION ON THE MASS MARKET
DEVELOPMENT FOR THE MAINSTREAM
ASSISTIVE TECHNOLOGIES
MARKETING AND DISTRIBUTION VIA HEALTH CARE PROVIDERS
DEVELOPMENT FOR PEOPLE WITH DISABILITIES

Fig. 1. Continuum of consumer and assistive technologies.

2 Related Work

There is currently not much research that specifically addresses people with ID and the use of user interfaces; studies available are predominantly related to specific applications or features, or specific types of disabilities. Research suggests that, to limit the digital divide and to "exploit the full potential of the respective device" [4], current digital technologies have to be made usable for people with ID through adaptation of their user interfaces or through guidance by non-disabled persons [4,6–8]. While simple or analog interfaces can be modified easily, e.g., adapting a door knob with clay or replacing a button with a bigger one, most natural user interfaces rely on pattern recognition, so they aren't as easy to be adapted [4].

2.1 Adaptation Possibilities for Interface Input Modalities

We differentiate between the input and output modality of the respective interface type. Since a suitable classification of input and output modalities of current interface types does not yet exist, we created our own. In the following sections, the different input and output modalities of the user interfaces included in the study are examined for their accessibility and adaptability at the current state of the art. For the sake of clarity and length of this study, it should be noted that this overview includes examples of adaptability and does not cover all possible solutions. Also, we focus on the input modality of the respective interface type. Other forms of input as the types stated below are *interfaces with multimodal input* - with more than one input modality, they can consist of two or more of the interface types described - and *interfaces with no input modality*. This is the case if the interface is triggered by someone or something else than the user or if there is only output.

Interfaces with Pointing Gestures: These are interfaces operated with two-dimensional pointing devices, like touch, computer mouse, or pencil input. Pointing gestures are among the most commonly used input modalities, especially touch interfaces have replaced other interface types in recent years and are nowadays used on smartphones, tablets or smartwatches as well as on desktop PCs or laptops [9]. Although Interfaces with pointing gestures, especially touch-sensitive interfaces are ubiquitous in society, people with ID still face problems in using them, caused by small screens, text or button sizes, difficult error handling, inadequate feedback, and a wide range of interaction methods [10,11]. Braun et al. analyzed different natural user interfaces and their accessibility for people with ID and found several (physical and cognitive) skills needed for using these types of interfaces. By comparing the needed skills with the actual skills of 44 people with ID, they found that 29.5% of the participants would face major problems in usage, 51.0% would still face minor problems and only 19.6% would be able to use the interface type without problems. Difficulties that the participants faced while using the interface were for example "small play button size, letting go of the button (pressing too long and using it as a physical button), keeping their whole hand on the screen and not being able to only use one finger of their hand to touch" [4].

Regarding system accessibility features (*category 1A* of the continuum, see Fig. 1) Apple developed several accessibility features for its touch interfaces, like "assistive touch" where the user can choose between different touch gestures (e.g., tap, double tap, hold) or create their own for certain actions; or "touch accommodations", where the user can change the hold duration (time until touch is recognized), tell the system to ignore accidental repeated taps and only recognizing one of them, or settings and assistance for swipe gestures. Also "predictive text" give suggestions for following words of a sentence [12]. Android also has certain accessibility features available, like "touch & hold delay" or "action blocks", especially addressing people with ID. It is possible to create different actions that can be triggered with only one touch gesture, making difficult processes with multiple steps easier to handle [13]. On Windows users can adjust and customize the color and size of the mouse cursor [14].

There are some applications available that can make interfaces with pointing gestures more accessible for people with ID. The "Shortcuts" app for iOS is comparable to Androids "Action Blocks" and users can create shortcuts of one or more actions [12]. Also, there are launcher apps (e.g., "BIG Launcher", "easierphone") available for the different operating systems that replace the home screen of the device (*category 1B*). Only the most important functions or apps (such as taking pictures or writing a message) can be stored there, which can lead to easier usage and less overwhelm caused by too much information [15,16]. Possible adaptations for touch interfaces can be finger guide grids, created with the help of a 3D printer, which can be adapted to the respective touch surface and application (*category 1C*). They provide security when selecting a field and simplify the selection [17]. For people that are not able to use this input modality correctly, it is also possible to adapt smartphones, tablets or PCs with alternate input methods, e.g., like additional buttons, pencils, controllers or external keyboards (*category 1C*) or using voice commands or head tracking instead of pointing gestures [18].

Interfaces with Buttons or Switch Elements: This input modality describes elements that provide two states, such as keyboards, joysticks, or buttons. Problems while using keyboards for people with ID or additional physical impairments can be the lack of writing skills [19] and therefore not understanding the different keys. Also, the complexity of the interface input can be a problem for people with ID. A study of 353 participants, ranging from mild to severe disability, found that 38.8% of the participants could not use a keyboard at all, and most of them performed poorly when using the keyboard [20]. More training, larger buttons or keyboard elements, or a simpler layout focusing on the most important elements could help improve usage. There are alternative keyboards and joysticks available for people with disabilities (*category 2*) with different sizes, layouts, or colors or the possibility to rearrange buttons [21]. Specifically for gaming, Microsoft has launched the Xbox Adaptive Controller, which connects to a range of devices and to which a variety of different buttons, switches, mounts, or joysticks can be connected to enable customized use [22]. For people with ID that can use keyboards but no other input modalities like touch, it is possible, for example, to use and navigate mobile devices like tablets with external keyboards. It is also possible to train keyboard shortcuts that perform certain tasks or use text replacements or suggested words (*category 1A*). Possible adaptations for keyboards can also be finger guide grids that help the selection and prevent the unintentional pressing of multiple keys (*category 1C*) [17]. People who are unable to use keyboards or lack literacy skills can alternatively use voice commands or dictate text [18].

Interfaces with Voice Interaction: Voice-controlled interfaces are triggered by voice or sound. They are most commonly used on specially developed devices such as Amazon Echo (Alexa) or Google Home (Google Assistant), but can also run on smartphones, tablets, or desktop PCs (e.g., Siri) [4]. They have the potential to be accessible and inclusive for people with ID and especially with limited mobility or blind users [23] and computerized speech interaction is even being used in therapy for people with speech impairments [24]. They can be useful for people with disabilities and help them to be more independent in their daily lives, for example by enabling them to operate a smart home [23,25], e.g., to regulate lights or temperature or play music and use the TV, or by using them for operating mobile devices and thus avoid spelling and typing problems [7]. However, accessibility challenges that people with ID face are primarily related to speech impairments or not using standard speech. Most speech recognition software requires clear pronunciation to recognize a command [7,23]. Balasuriya et al. propose adjustable input settings for voice assistants to be more accessible for people with pronunciations that vary from the norm [7].

There are some operating system features (*category 1A*) for voice input devices that improve accessibility, however, most of them refer to the speech output of the device, not to the user's speech input. For Amazon Alexa it is possible to change the input modality and use the device with touch instead of voice [26]. Google Nest allows the user to set a start and end sound when the Assistant has recognized the wake word ("Hey Google", "Ok, Google") and has processed the voice command [27]. While these settings do not address the issues of people with ID and speech recognition, there are solutions that do address them: "Voiceitt" (*category 1B*) is an app specially developed for people with non-standard speech to use voice assistants. By training different commands or statements, the app learns the user's individual speech pattern. Any kind of repeatable audio pattern can be trained, no matter the speech impairment. When the user gives a trained command, it is translated into the desired expression and by linking it to Alexa, e.g., smart home devices can be controlled. There is also a communication mode in which unclear speech is translated into words and sentences that can be understood by others. Through machine learning and statistical modeling, "Voiceitt" improves with each use. The developers of "Voiceitt" are also working on an expansion that recognizes spontaneous speech and doesn't require training [28].

Interfaces with Object Interaction: These are interfaces that connect the digital and physical world by manipulating real objects to trigger an action. They are controlled via a 3D interaction element and rely on the sense of touch [29]. Although this input modality is rarely used in consumer technologies at the moment, it could have great potential for people with ID. One example of a consumer technology that uses object interaction is the "Toniebox" (category 1) - an audio system that is controlled by putting little figures on top of a box to choose songs or stories to be played [30]. While the official "Toniebox" cannot easily be adapted, the modified "Phoniebox" - a project of the maker community - can (*category 1C*). There is a large community of makers who are constantly developing and improving the "Phoniebox". The interface is fully customizable, but since it is a "DIY" project, it must be built by the users themselves, which requires programming and various making skills [31]. The objects needed for interaction can be individually selected and customized, they just need to be linked to an RFID tag, which can possibly have a positive effect on the accessibility of this input modality. People with ID or their caregivers and relatives can thus choose objects that are usable and tangible for the individual person.

Interfaces with Touchless Interaction: Interfaces with touchless input are controlled via 3D body or hand gestures and are tracked via depth sensors. They allow the user to trigger actions without physically touching an interaction device like a keyboard or screen. They have not fully entered the mainstream as an input modality of consumer technologies yet and are mostly used in gaming applications (e.g., with the Microsoft Kinect), virtual reality (e.g., with a

Leap Motion Controller) or semi-public places (e.g., in exhibitions or museums) and therefore few applications with touchless input have been made available to people with ID [4,32]. Touchless approaches in the field of assistive technologies exist that enable the control of wheelchairs or other aids (*category 2*). They are particularly interesting for people with restrictions in their motor control who cannot use buttons, joysticks, or touch switches [33]. The capabilities of touchless consumer devices are constantly improving, e.g., the Leap Motion Controller can detect hands and corresponding fingers including small movements. This interaction method could also be relevant for people with disabilities (and people with ID), however, these users often do not have the necessary skills, such as interaction precision or speed, to use this type of interface, which makes it mandatory to adapt and personalize the interaction [32]. An application with the Leap Motion Controller and people with ID was tested by Braun et al. [4]. The authors found that this interface was difficult to use for people with ID who had an additional motor disability or had problems with fine motor skills. Also, the concept of touchless interaction was difficult to understand for many of the users, as evidenced by them touching the device instead of performing touchless gestures. Although the Leap Motion Software currently only supports a small number of gestures, it is possible to train custom gestures [34], which could be beneficial for people with ID.

2.2 Adaptation Possibilities for Interface Output Modalities

Since interfaces always have an input and output modality, possible output modalities are briefly described here for completeness but are not the main focus of this study. If interface feedback consists of more than one output modality it is classified as multimodal output.

Interfaces with Visual Feedback: This covers all forms of visual feedback, like text, motion graphics, pictures, LEDs or colors. A study in 2013 with over 1600 students with ID found that "29.3% do not read at all, 6.8% read at a logographic stage, 31.9% at an alphabetic and 32% at an orthographic level" [19]. Problems that can arise for people with ID can be small text or button sizes [4,10,11], and also lack of literacy skills. Additional disabilities like impaired vision or blindness can also affect the usage of this output modality [4]. Apple has several built-in accessibility features regarding visual feedback (*category 1A*) that can help people with ID use interfaces with visual feedback better. It is possible to adapt font size and strength or use display zoom, which enlarges everything. Also, light or dark modes can be used, and it is possible to increase the contrast or invert colors or adjust screen brightness. For lesser distractions and sensory overload, it is possible to cancel out ads or navigation bars when using safari (with Safari Reader). The size of app icons on the home screen can be changed or used to only display important apps, and the movement of onscreen elements can be decreased, also leading to less sensory overload. People with a lack of literacy skills or vision problems can use "VoiceOver", which is a screen reader that describes elements on the screen or “Spoken Content”, which also reads out the content of the screen [18].

Interfaces with Auditive Feedback: These interfaces use sound or voice notifications as output. Problems that could arise for people with ID are not understanding the auditive output - e.g., because of hearing loss, the output speed or possibly distracting sounds or sound effects [23]. System accessibility features that are currently available for Amazon Alexa are settings to change the preferred speech rate and adjustable volume for timers, alarms or media (*category 1A*). Amazon Echo devices can be paired with certain Bluetooth speakers for better sound and some can display captions (visually) [26].

Interfaces with Haptic Feedback: This covers all sorts of haptic feedback which is mostly vibrations or tactile feedback. For people with ID, an interface with haptic feedback can “potentially contribute to an enhancement in perception of objects and overall ability to perform manipulation tasks” [35]. For mobile devices like smartphones, tablets or smart watches haptic feedback has already been established, like for notifications, phone calls, or alarms. These vibrations can be turned off and on and the intensity can be adjusted [12] (*category 1A*). Tactile feedback is also used often in navigation systems to deliver navigation information, increase situation awareness, and to support eyes-free usage (*category 1B*) [36]. There are several wearables available for navigation, e.g., vibration belts (*category 2*) [37], which could also possibly help people with ID make navigating easier.

3 Methodology

To find out which consumer technologies, interface input modalities and their adaptations are currently usable and suitable for people with ID, a study was performed in which people with ID were asked about their most important participation wishes to improve their everyday life through consumer technology. These wishes were analyzed by the researchers according to feasibility and cost-effectiveness to find suitable consumer solutions.

The individual abilities of people with ID were considered when selecting the technology and the respective interface type. The prototypical solutions, often several solutions with different input and output modalities and different degrees of adaptation, were tried out with the participants in a technology testing scenario. Together with participants and their caregivers and relatives we were then able to decide in a participatory manner which solution was the most appropriate for the participant. These solutions have been evaluated and analyzed for their input modality and adaptation level in the presented study.

3.1 Target Group

This study includes people with various degrees of ID, some with additional motor impairments. Based on this, 3 target groups were identified:

1. *Individuals with mild ID*, who can speak and, if applicable, read and write (with motor limitations, if applicable).
2. *Persons with moderate ID*, who can understand simple language and can express themselves with limited speech (if applicable, with motor limitations).
3. *Persons with multiple disabilities* in the sense of ID with severely impaired intentionality and understanding of symbols combined with significant motor impairments.

Access to the field has been secured through three different institutions from the disability sector in southern Germany, where people with ID live in residential or outpatient facilities. Recruiting this target group can be a time-consuming and difficult process since some people with ID cannot give their consent to participate in studies themselves and parents or legal advisors must give their consent instead. To best represent the interests of all participants, an ethical application was approved by the German Society for Educational Science (DGfE) and data was anonymized.

3.2 Case Selection and Planning

The following method was chosen for case selection and planning of the possible technical solutions, matching the identification of participation wishes:

1. *Case selection by researchers according to the following criteria:*
   - expected improvement in participation
   - technical feasibility
2. *Technology identification:* Workshops were held with the researchers and 4 experts in aided communication, participation, assistive technology, human-machine interaction, and interface/interaction design. The goal was to find out which participation wishes were feasible and to collect 1–3 solution ideas per case, matching the cognitive and physical abilities of the participants, in order to use them for further development of prototypes. Consideration was given to how the solution could be implemented technically and how much effort would be required. The workshops consisted of a presentation of existing solutions (based on previous research), followed by an initial phase in which each participant considered solutions individually. Subsequently, solution ideas were developed through discussions with all workshop participants and the further procedure for the development of the prototypes was determined.

3.3 Participatory Technology Testing and Selection

In order to find the most suitable solution, 1–3 ideas were presented to the participants and then evaluated and discussed in a participatory manner. This process was based on the user-centered approach of *Scenario-Based Design* [38]. It varied depending on the type of solution and individual abilities, e.g., visual or written solution scenarios, prototype testing, wizard-of-oz testing, or actual technical solutions if they were already implementable (e.g., installing a certain app, for example for using smart home devices with non-standard speech like “Voiceitt” or setting up accessibility features like bigger fonts or screen readers for easier interaction with a device). Here, the participants were given simple tasks to solve (e.g.: “Turn on the music using the Voiceitt app” or “Open the Deutsche Bahn app and enter your destination in the search bar”). Afterwards, the pros and cons of the different solutions and possible adaptations were discussed with the participants and/or their caregivers. This resulted in one or more (depending on the number of realizable wishes) suitable solutions for each of the participants.

4 Study

Here we describe our user study, which took place from June 2021 to July 2022. 43 people with ID who had also previously attended the identification of participation wishes participated. This resulted in 150 wishes. The number ranged from 1 to 7 wishes per person, with a mean of 3.5 wishes. The participation wishes surveyed primarily concerned more independence in everyday life in various areas such as entertainment, mobility/navigation, household tasks (e.g. shopping), learning, or (digital) communication. The IDs ranged from mild intellectual disabilities to more serious disabilities and were often combined with motor impairments. Of the 43 participants, 18 were categorized in target group 1, 22 in target group 2, and 3 in target group 3. Of the 18 participants in target group 1, 4 had an additional motor impairment and one had an additional sensory impairment. In target group 2, 7 had an additional motor impairment. All the people in target group 3 had severe additional motor impairments. The age of the participants ranged from 24 to 76 years, with an average of 48.8 years. Of the 43 participants, 29 identified as male and 14 as female.

4.1 Study Setup

In addition to the participant, one or two staff members from the research team participated in the technology selection. In many cases, a close relative or caregiver was also present. The appointments took place in inpatient residential facilities within these, in familiar surroundings. For evaluation, video recordings with two cameras from two angles, an additional sound recording, and observation protocols were made.

4.2 Case Selection and Planning and Participatory Technology Testing and Selection

The case selection and planning resulted in various proposed solutions, possibly suitable for the cognitive and physical abilities of the participants. The participatory technology testing resulted in the selection of different types and technologies and interfaces. In this process, 34 participation wishes were eliminated. The reasons for this were, for example, lack of feasibility, no added value provided by a technical solution, or no further participation of the person in the study due to personal reasons. The 116 possible solutions - belonging to the remaining 41 participants - will be examined in this study. For the analysis, each solution is classified into the interface input and output modalities introduced in Sect. 2, as well as into the adaptation level category based on the proposed continuum.

5 Findings

In this section, the collected data is analyzed, and our findings are presented. As stated before, 116 technical solutions for 41 people with ID were examined.

5.1 Interface Input Modality and Level of Adaptation

Looking at Table 1 and 2, we can analyze the suitable interface input types for the participants. 67 consumer solutions had one input modality, 41 used multimodal input with two input modalities, and 3 solutions used three input modalities. 5 solutions had no input modality. The level of adaptation (Table 3) refers to our classification in the *continuum of consumer and assistive technologies* introduced before. The percentage in brackets refers to the total occurrence of the respective input modality. Since we focused on consumer technologies for possible solutions, *category 2 (special assistive devices)* is not applicable and therefore not evaluated.

Table 1. Interface Input Modalities.

Input Modality (N)	Overall	N=1	N=2	N=3
Pointing gestures	95	60 (89.6%)	32 (68.1%)	3 (100.0%)
Voice interaction	25	0	23 (48.9%)	2 (66.7%)
Buttons or switch elements	22	5 (7.5%)	12 (25.5%)	3 (100.0%)
Object interaction	17	1 (1.5%)	15 (31.9%)	1 (33.3%)
No input modality	5	0	0	0
Touchless interaction	1	1 (1.5%)	0	0
Total	116	67	41	3

Interfaces with pointing gestures are, unsurprisingly, the most used interface type, appearing 95 times in the 116 solutions. This makes sense since this type of input modality is used in most devices nowadays [9]. It is used 60 times as a single input modality, 32 time alongside one other input modality (23 times with “voice interaction”, 6 times with “object interaction” and 3 times with “buttons or switch elements”) and 4 times alongside 2 other input modalities (two times

Table 2. Multimodal Input Modalities.

Input Modalities	N
Pointing gestures + Voice interaction	23 (19.8%)
Object interaction + Buttons or switch elements	9 (7.8%)
Object interaction + Pointing gestures	6 (5.2%)
Pointing gestures + Buttons or switch elements	3 (2.6%)
Pointing gestures + Buttons or switch elements + Voice interaction	2 (2.6%)
Object interaction + Pointing gestures + Buttons or switch elements	1 (0.9%)

Table 3. Level of Adaptation.

Input Modality	*Category 1*	*Category 1A*	*Category 1B*	*Category 1C*
Pointing gestures	26 (43.3%)	4 (6.7%)	30 (50.0%)	0
Voice interaction	0	0	0	0
Buttons or switch elements	1 (20.0%)	0	4 (80.0%)	0
Object interaction	0	0	0	1 (100%)
Touchless interaction	1 (100.0%)	0	0	0
Pointing gestures + Voice interaction	7 (30.4%)	4 (17.4%)	12 (52.2%)	0
Object interaction + Buttons or switch elements	0	0	0	9 (100.0%)
Object interaction + Pointing gestures	4 (66.7%)	1 (16.7%)	0	1 (16.7%)
Pointing gestures + Buttons or switch elements	2 (66.7%)	0	1 (33.3%)	0
Pointing gestures + Buttons or switch elements + Voice interaction	0	0	0	2 (100.0%)
Object interaction + Pointing gestures + Buttons or switch elements	1 (100.0%)	0	0	0

with “buttons or switch elements” and “voice interaction” and one time with “object interaction” and “buttons or switch elements”). Because these types of input modalities are already widely used in current consumer technologies and have more accessibility features than other interface types, they can already be used as they are by a lot of participants with ID (43.3% in *category 1*, where no adaptation has occurred and 6.7% has been adapted with accessibility features (*category 1B*). However, in 50.0% of the cases where this input modality was used, it had to be adapted with assistive applications to be usable for the participants, which highlights the demand for assistive apps for this interface input type for people with ID.

Interfaces with voice interaction are the second most used input modality with 25 occurrences in total. Interestingly, they are never used as a single input modality, only alongside other types of input. They are used 23 times with "pointing gestures" and 2 times in combination with two other input modalities (with "pointing gestures" and "buttons or switch elements"). This interface type could not be used as a standalone input modality in our study, which shows that voice-only solutions bear difficulties for this target group. Most of the time, they are used in combination with interfaces with pointing gestures, where 30.4% belong to *category 1*, 17.4% belong to *category 1A* and 52.2% belong to *category 1B*. When trying out the different solutions with people with ID in the study, it also often became apparent how little speech input currently works for people with different speech impairments and how poorly they are often understood by the technology. This usually led to frustration and is the reason that in the end other input modalities were selected for the participant or a multimodal input was chosen. Voice interfaces need more adaptation options for people with disabilities so that people with all types of speech can use this input modality. Solutions like *Voiceitt* [28] have already taken a big step in this direction.

Interfaces with physical buttons or switch elements are the third most used interface input type in the study, with 22 uses in total, 5 times as a single input modality, 14 times alongside one other input modality (9 times with "object interaction" and 3 times with "pointing gestures"), and 3 times alongside two other input modalities (2 times with "pointing gestures" and "voice interaction" and 1 time with "object interaction" and "pointing gestures"). As a standalone input type, 80.0% belonged to *category 1B*, and 20.0% to *category 1*. Used with object interaction, all solutions belonged to *category 1C*, in combination with pointing gestures, 66.7% categorized in *category 1* and 33.3% in *category 1B*.

Interfaces with object interaction are used 17 times in total. This shows the potential for certain people with ID, even though this input modality is not often used in consumer technologies. It is only once used as a standalone input modality and 15 times alongside one other input type (9 times with "buttons or switch elements" and 6 times with "pointing gestures"). It is used one time with two other input modalities ("pointing gestures" and "buttons and switch elements"). This input type was only used once as a single input modality in *category 1C* (where hardware or sensory adaptions had been made) and was most of the time (n = 9) used in combination with "buttons or switch elements", where also all solutions belonged to *category 1C*. In combination with "pointing gestures" (n = 6) adaptation wasn't needed as often (66.7% of the times belonging to *category 1*, 16.7% *category 1A* and 16.7% *category 1C*. The numbers imply that interfaces with object interaction must be adapted more when they are used alone or in combination with "buttons or switch elements" and less when they are combined with "pointing gestures", as there are already more adaptation possibilities available for this interface type.

Interfaces with touchless interaction were only used once, as a standalone input modality without adaptation. This supports the statement that this type of interface has not yet finally arrived in consumer technologies [4,32] and that it is therefore more difficult to use the available solutions with people with ID. Before this type of input can be used in consumer technologies by people with ID it has to become standard for the regular user and possible adaptation possibilities must be developed. However, the use of this input type in various assistive technologies shows the potential for certain types of disabilities [33].

5.2 Interface Output Modality

In total, 105 interfaces used visual feedback (44 as a standalone output modality), 72 used audio feedback (11 as a standalone output modality) and 6 used haptic feedback (none as a standalone output modality). 67 solutions used multimodal feedback, the most used multimodal feedback was audio and visual (n=56). Haptic feedback was always used in combination with visual feedback.

5.3 Correlation Between Adaptation Level, Target Group and Age

We investigated the correlation between the *Target Group* (Group 1, 2 or 3) and *Level of Adaptation* (*categories 1-1C*) and *Age* using Pearson's correlation. This showed a positive relationship of r=0.282 (p-value=0.002) and means that *Target Group* has a weak effect on *Level of Adaptation* of possible consumer technologies and people with more serious IDs or multiple disabilities possibly need more adaptation. Also, *Age* had a weak positive relationship with *Level of Adaptation* with r=0.316 (p-value=<0.001) showing that age could play a factor in the needed adaptation level of consumer technologies.

5.4 Target Group and Input Modality

The most used interface input type in target group 1 was pointing gestures (47.3%) and pointing gestures + voice interaction (30.9%). In group 2 it was also pointing gestures (58.6%), pointing gestures + voice interaction (10.3%), and object interaction + buttons and switch elements (10.3%). In group 3 the only interface input types used were object interaction + buttons or switch elements (66.7%) and interfaces with no input modality (33.3%). This shows that especially for people with more severe disabilities, interfaces with physical object interaction + buttons and switch elements may have a big potential for participation and that voice interaction now can mostly be used by people with less severe disabilities.

6 Discussion and Conclusion

This study focused on current user interfaces, especially their input modalities. They were analyzed on suitability and adaptation options for people with ID. The requirements that this target group has for user interfaces and the adaptations that need to be made to make them usable were evaluated. Our user study resulted in 116 possible solutions and prototypes with different input modalities and varying degrees of adaptation, tested by 41 participants with ID, with the aim of being integrated into the participants' everyday lives. We found, that interfaces with pointing gestures (e.g., touch or mouse input) are currently the preferred interface type for most people with ID, as this input type is used in most technologies today and provides the most accessibility features. Also, voice input has a high potential for a lot of people with ID but must currently be used as a multimodal input type, alongside one or more input modalities, as many necessary accessibility features are missing. As voice input continues to improve and adapt to the needs of people with ID, this input modality could be used by many more. Especially people with more severe and multiple disabilities (target group 3) are unable to use most interface types. The level of disability (target group) had a significant positive correlation to the level of adaptation needed, which shows that people with more severe disabilities need more adaptation possibilities for current consumer technologies and interfaces. Also, age of the participants plays a factor in the level of adaptation needed.

The analysis of related work and state of the art has also shown that currently, most accessibility features and possible adaptations are developed for interface input with pointing gestures (like touch or mouse input). Other input types, like voice interaction or object interaction, may have a big potential for people ID, but are currently more difficult to adapt to the individual needs of different users. There are approaches to make more interface types accessible to this target group, but in the future, the needs and abilities of people with disabilities, especially people with ID, need to be much more involved in the design and development process of new technologies.

There are some limitations to our research. Since there are no well-researched methods for conducting usability studies or evaluating different interfaces with people with ID, an exploratory approach was chosen. Also, not every interface input type could be tested with every participant, as the consumer solutions depended heavily on each person's participation wishes and abilities. This causes, for example, that not all input modalities occur equally often, which makes the comparison more difficult (e.g., touchless interfaces only occurred one time). Our study also did not yet include how training and longer-term use can affect the usability of an interface type. In future work, we will investigate the usability and suitability of the different solutions and interfaces in the everyday life of the participants. It will be shown how useful the individual solutions are and whether they can be integrated into their daily lives and are accepted in the long term. This study highlights the current potential of interface used by people with ID and emphasizes the need for more adaptability of consumer technologies.

Acknowledgements. This study is part of a project funded by the Federal Ministry of Education and Research (BMBF) in Germany within the framework of the program "FH Sozial 2017". We would also like to thank the participating institutions for people with disabilities.

References

1. Dufva, T., Dufva, M.: Grasping the future of the digital society. Futures **107**, 17–28 (2019)
2. Lussier-Desrochers, D., et al.: Bridging the digital divide for people with intellectual disability. Cyberpsychology: J. Psychosoc. Res. Cyberspace **11**(1), 1 (2017)
3. European Parliament and Council. Directive (EU) 2019/ of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services. Official Journal of the European Union, p. 46, April 2019. https://eur-lex.europa.eu/eli/dir/2019/882/oj
4. Braun, M., Wölfel, M., Renner, G., Menschik, C.: Accessibility of different natural user interfaces for people with intellectual disabilities. In: 2020 International Conference on Cyberworlds (CW), pp. 211–218. IEEE, Caen, France (2020)
5. Congress gov. S.2561 - 100th Congress (1987–1988): Technology-Related Assistance for Individuals With Disabilities Act of 1988 (1988). https://www.congress.gov/bill/100th-congress/senate-bill/2561
6. Jimenez, B.A., Alamer, K.: Using graduated guidance to teach iPad accessibility skills to high school students with severe intellectual disabilities. J. Spec. Educ. Technol. **33**(4), 237–246 (2018)
7. Balasuriya, S.S., Sitbon, L., Bayor, A.A., Hoogstrate, M., Brereton, M.: Use of voice activated interfaces by people with intellectual disability. In: Proceedings of the 30th Australian Conference on Computer-Human Interaction, pp. 102–112. ACM, Melbourne, Australia (2018)
8. Barlott, T., Aplin, T., Catchpole, E., Kranz, R., Le Goullon, D., Toivanen, A., et al.: Connectedness and ICT: opening the door to possibilities for people with intellectual disabilities. J. Intellect. Disabil. **24**(4), 1–19 (2019)
9. Gündogdu, R., Bejan, A., Kunze, C., Wölfel, M.: Activating people with dementia using natural user interface interaction on a surface computer. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare - PervasiveHealth 2017, pp. 386–394. ACM Press, Barcelona, Spain (2017)
10. de Urturi Breton, Z.S., Jorge Hernandez, F., Mendez Zorrilla, A., Garcia Zapirain, B.: Mobile communication for intellectually challenged people: a proposed set of requirements for interface design on touch screen devices. Commun. Mob. Comput. **1**(1), 1 (2012)
11. Williams, P., Shekhar, S.: People with learning disabilities and smartphones: testing the usability of a touch-screen interface. Educ. Sci. **9**(4), 263 (2019)
12. Apple. Official Apple Support (2022). https://support.apple.com/
13. Google. Android Accessibility Help (2022). https://support.google.com/accessibility/android
14. Microsoft. Windows Accessibility Features — Microsoft Accessibility (2022). https://www.microsoft.com/en-us/accessibility/windows
15. 2BIG s r o. BIG Launcher. https://biglauncher.com/en/
16. Pappy GmbH. Easierphone. https://easierphone.com/

17. RehaMedia. Fingerführraster. https://rehamedia.de/glossar-lexikon/fingerfuehrraster/
18. Apple. Accessibility (2022). https://www.apple.com/accessibility/
19. Ratz, C., Lenhard, W.: Reading skills among students with intellectual disabilities. Res. Dev. Disabil. **34**(5), 1740–1748 (2013)
20. Li-Tsang, C., Yeung, S., Chan, C., Hui-Chan, C.: Factors affecting people with intellectual disabilities in learning to use computer technology. Int. J. Rehabil. Res. **28**(2), 127–133 (2005)
21. SG Enable. Assistive Technology - Disability Support | Enabling Guide (2022). https://www.enablingguide.sg/im-looking-for-disability-support/assistive-technology/at-intellectual-disability
22. Microsoft. Xbox Adaptive Controller | Xbox (2022). https://www.xbox.com/en-US/accessories/controllers/xbox-adaptive-controller
23. Pradhan, A., Mehta, K., Findlater, L.: Accessibility came by accident: use of voice-controlled intelligent personal assistants by people with disabilities. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI 2018, pp. 1–13. ACM Press, Montreal QC, Canada (2018)
24. Palmer, R., Enderby, P., Hawley, M.: Addressing the needs of speakers with long-standing dysarthria: computerized and traditional therapy compared. Int. J. Lang. Commun. Disord. **42**(s1), 61–79 (2007)
25. Domingo, M.C.: An overview of the internet of things for people with disabilities. J. Netw. Comput. Appl. **35**(2), 584–596 (2012)
26. Amazon. Accessibility Features for Alexa (2022). https://www.amazon.com/gp/help/customer/display.html?nodeId=202158280
27. Google. Google Nest display accessibility settings - Android - Google Nest Help (2022). https://support.google.com/googlenest/
28. Voiceitt Inc., Voiceitt. https://voiceitt.com/
29. Shaer, O.: Tangible user interfaces: past, present, and future directions. Found. Trends Hum. Comput. Interact. **3**(1–2), 1–137 (2009)
30. tonies GmbH. tonies (2022). https://tonies.com/en-gb/
31. Flor M. Phoniebox: the RPi-Jukebox-RFID; 2022. Original-date: 2017–02-02T11:41:41Z. https://github.com/MiczFlor/RPi-Jukebox-RFID
32. Augstein, M., Kurschl, W.: Modelling touchless interaction for people with special needs. In: Koch, M., Butz, A., Schlichter, J.H., (eds.) Mensch & Computer 2014 - Workshopband. De Gruyter Oldenbourg , Berlin (2014). Accepted: 2017–11-22T15:08:52Z. http://dl.gi.de/handle/20.500.12116/8164
33. Kouroupetroglou, G., Das, P.: Assistive Technologies and Computer Access for Motor Disabilities: Advances in Medical Technologies and Clinical Practice. IGI Global, USA (2014)
34. Jamaludin, N.A.N., Huey, O.: Dynamic hand gesture to text using leap motion. Int. J. Adv. Comput. Sci. Appl. **10**(11), 199–204 (2019)
35. Jafari, N., Adams, K.D., Tavakoli, M.: Haptics to improve task performance in people with disabilities: a review of previous studies and a guide to future research with children with disabilities. J. Rehabil. Assist. Technol. Eng. **3**, 205566831666814 (2016)
36. Pielot, M., Poppinga, B., Heuten, W., Boll, S.: PocketNavigator: studying tactile navigation systems in-situ. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3131–3140. ACM, Austin, Texas, USA (2012)
37. feelSpace GmbH. naviGürtel (2022). https://feelspace.de/
38. Carroll, J.M., Rosson, M.B., Farooq, U., Xiao, L.: Beyond being aware. Inf. Organ. **19**(3), 162–185 (2009)

A Rule Mining and Bayesian Network Analysis to Explore the Link Between Depression and Digital Behavioral Markers of Games App Usage

Md. Sabbir Ahmed[1](✉), Tanvir Hasan[1], Md. Mahfuzur Rahman[2], and Nova Ahmed[1]

[1] Design Inclusion and Access Lab (DIAL), North South University, Dhaka 1229, Bangladesh
msg2sabbir@gmail.com, tanvirfuad00@gmail.com,
nova.ahmed@northsouth.edu
[2] Department of Computer Science and Engineering, Eastern University, Dhaka 1345, Bangladesh
dean_et@easternuni.edu.bd

Abstract. Amid the COVID-19 pandemic, spending time on Games increased much, which may impact mental health. While numerous studies were conducted exploring the relation between Games and depression, none of the studies used objective (i.e., actual) Games app usage data which could provide unbiased and real-time insights. To fill this research gap, using our developed app that retrieves the past 7 days' actual app usage data accurately, we conducted a study on Games app users ($N = 60$) in Bangladesh. We extracted the behavioral markers from the foreground and background Games app usage events' data. To explore the relation between Games and depression, we mined rules, did correlation analysis, and built Bayesian networks. Our analyses demonstrated that the students who spent higher time and had a higher launch per Games app on weekends were more likely to be depressed ($p < .05$). In addition, from the Bayesian analysis, we found that while some usage data impacts depression, depression also impacts some usage behavior such as frequency of launching Games apps. Apart from raising awareness about the negative impact of Games, insights from our study can facilitate the design of systems to improve the students' mental health.

Keywords: Smartphone · Games · Depression · Behavioral patterns · Bayesian network

1 Introduction

Amid the pandemic, Games playing time increased and 75% of the rise is estimated to persist in the next two years [11]. However, problematic gaming is found to have a negative impact on mental health [24, 29] which shows the necessity of in-depth exploration of the link between Games and psychological problems. In Bangladesh, the rate of psychological problem depression is higher among university students compared

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 557–569, 2023.
https://doi.org/10.1007/978-3-031-34586-9_37

to other groups of people [10]. The stay at home for the pandemic and the high availability of smartphones where 86% of Bangladeshi university students have smartphones [2] can facilitate them in increasing Games playing.

With great interest, scholarly articles explored the relation between Games and depression where most studies (e.g., [5, 24, 28]) used subjective data. Some of these studies found a positive association of depression with addiction to video gaming [5] and problematic online gaming [24]. However, as subjective data does not present the actual app usage behavior [18, 32], the findings of these studies may not unveil the exact relation.

On the other hand, previous studies also explored objective data on gaming as well as of app categories. In the case of gaming, researchers explored gaming data for purposes such as exploring the feasibility of gamification in having a positive impact on sleep-wake [12], exploring patterns in Games playing behavior [8], to distinguish the gaming events from the sensor-collected data [30], to assess problematic internet use [31], and to find out the factors related with underreported playing time [32]. In a recent study [18], researchers used both subjective and objective data from an online chess platform to explore the relationship with problematic effects (e.g., disrupted sleep). Objective behavioral markers of different app categories such as Communication [1, 15], Health & Fitness, Photo & Video [1], and Social Media [1, 15, 27] have also been explored in different contexts. Researchers [15] observed that depressed and non-depressed students have significantly different app usage duration and frequency of launching Communication apps. In addition, depressed students have significantly higher unique app signatures in the case of Social Media, Health & Fitness app categories [1]. Though some studies explored the behavioral patterns of the depressed [1, 15], Hunt et al. [27] did a causal analysis by keeping students in the control and experimental groups. Their analysis demonstrated that limiting Facebook, Instagram, and Snapchat use reduce depression [27]. Objective behavioral markers regardless of the app categories have also been explored in a recent study [33] to classify the depressed and non-depressed through computational models. However, as far as we know, no study used objective Games app usage data to explore the relation between depression and Games apps usage, although Games is the most popular app category in Google Play Store [6] and amid the pandemic, Games playing time increased significantly [11].

Therefore, we explore the link between objectively measured Games app usage data and depression (i.e., Patient Health Questionnaire-8 (PHQ-8) scale's score [14]) and contribute to the pervasive health research area in the following ways.

- Firstly, to our best knowledge, using objective data, this is the first study to explore the relation between a psychological problem and digital behavioral markers of Games app usage which can provide unbiased findings.
- Secondly, through rule mining, we present Games app behavioral patterns that are associated with depression which can be potential to understand the nuanced differences between depressed and non-depressed students. In addition, this can facilitate in the development of computational models to predict depression leveraging digital behavioral markers.

- Thirdly, we develop a Bayesian network that shows that all behavioral markers of Games app usage do not impact depression and vice versa which can be useful to design pervasive systems for intervention.

2 Methods

2.1 Participants and Research Ethics

Our study was approved by a university from Bangladesh. We did the study in 2020 during the COVID-19 pandemic where 100 Bangladeshi students from 12 higher educational institutes participated. Among them, 60 students used the Games apps (please, see Sect. 2.2.2 for details) on which we conducted this study. All participants' data were collected through their consent and in the consent form, we specifically mentioned the required permission, collected data, data security, usage of their data, etc. Apart from this, to make the participants more aware of the collected data, our app asked for permission in runtime to access the app usage data before retrieving it.

2.2 Tools and Analysis

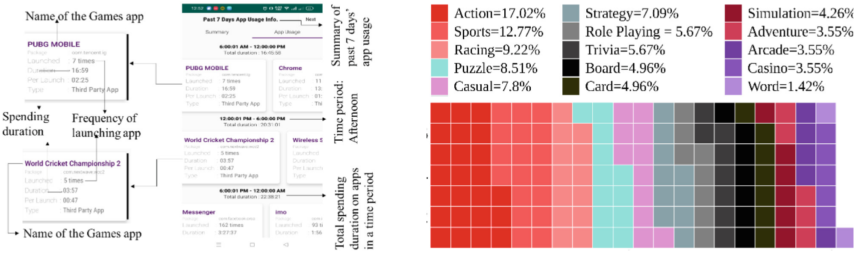
2.2.1 Depression Measurement

To measure the participants' depression, we used the 8-itemed PHQ-8 scale's score [14]. We explored the PHQ-8 scores as the continuous values in the correlation and Bayesian network analysis as described in Sect. 2.2.4 and Sect. 2.2.5 respectively. In addition, to understand the participants' depression and also as a requirement in the rule mining through the classification-based association algorithm (please, see Sect. 2.2.3), we divided the participants based on depression score. In finding major depressive disorder, the sensitivity and specificity are 100% and 95% respectively for a PHQ-8 score of 10 [14]. Hence, the participants who had a PHQ-8 score of 10 or more were grouped as the depressed and others (PHQ-8 $<$ 10) as the non-depressed participants.

2.2.2 Data Collection Tool and Extraction of Games App Usage Markers

As subjective data does not reflect the actual habit [18, 32], we retrieved participants' actual app usage data through our developed Android app (Fig. 1(a)). We used the Java class *UsageStatsManager* [9] to retrieve foreground and background events' data, and to store it, we used the Google Firebase database. We tested the app on 9 phones, and also compared it with the retrieved app usage data of the available such apps (e.g., [17]) in the Play Store. Since the system keeps app usage events' data only for a few days [9], our app can collect the past 7 days' app usage data very accurately.

Our developed app retrieved 817,404 foreground and background events data from 100 participants who used 1,129 apps. Two researchers and one student categorized these apps following the app categories of the Play Store, the app categorization process of previous studies (e.g., [1]), and an understanding of the apps' features. Among 1,129 apps, we found 141 (12.49%) apps in the Games category which consisted of 40,339 foreground and background events. The used Games apps were of 15 subcategories (Fig. 1(b)). Most (17.02%) apps were of Action (e.g., PUBG MOBILE) and the least



PUBG MOBILE
Package : com.tencent.ig
Launched : 7 times
Duration : 16:59
Per Launch : 02:25
Type : Third Party App

World Cricket Championship 2
Package : com.nextwave.wcc2
Launched : 5 times
Duration : 03:57
Per Launch : 00:47
Type : Third Party App

6:00:01 AM - 12:00:00 PM
Total duration : 16:45:58

12:00:01 PM - 6:00:00 PM
Total duration : 20:31:01

6:00:01 PM - 12:00:00 AM
Total duration : 22:38:21

Messenger
Launched : 162 times
Duration : 3:27:37

(a) Screenshot of the data collection tool showing Games app usage

(b) Number of apps in each of the 15 subcategories

Fig. 1. Data collection tool and participants' Games app usage.

(1.42%) were of Word (e.g., Word Forest) subcategory. On the other hand, among the 100 participants, 60% (N = 60) participants were Games app users who launched at least one Games app in 7 days. Since having the value 0 for the remaining 40% of non-users can make the variables' data distribution highly skewed, their data were excluded from the analysis to have unbiased findings.

From the retrieved app usage events, we quantified each participant's behavioral markers of Games app usage by calculating spending duration, frequency of launching apps, number of used apps, duration per app, duration per app launch, and frequency of launch per app of the Games category. In addition, to explore the Games app usage pattern, we calculated entropy $E(j) = \sum_{i=0}^{n} p(i) log p(i)$ where $p(i)$ indicates the probability to use the i^{th} Games app by the j^{th} participant. Since the app usage behavior varies by weekdays and weekends [1], to understand the association of PHQ-8 score with Games app usage data, we explore these 7 variables by calculating each variable's data of weekends (Friday and Saturday), weekdays (Sunday to Thursday), and 7 days (whole week). We divide the days based on the working week in Bangladesh.

2.2.3 Rule Mining and Extraction of Behavioral Patterns

Each of the aforementioned 7 variables presents Games app usage behavior and we denote this set of behavioral items by T_j for j^{th} participant. Using classification-based association (CBA) algorithm [16], we mine the behavioral patterns that are associated with depression in the form of $A_j \rightarrow B_j$ where A_j is the rule body containing a subset of T_j items and B_j is the rule head denoting the class (e.g., depressed) of j^{th} participant. Since CBA works with discrete values, inspired by the previous studies (e.g., [23]), we discretized the values of each behavioral marker into three equal groups where top one-third and bottom one-third percentile were grouped as the high and low users respectively and others were grouped as the medium users. For each rule, there are three parameters namely support, confidence, and lift based on which a rule is selected. Support ($\frac{frequency(rule\,body, rule\,head)}{N}$) denotes the frequency of a set of items appearing among all participants (N) whereas confidence ($\frac{frequency(rule\,body, rule\,head)}{frequency(rule\,body)}$) says how likely the rule head is to occur when the rule body appears. Having lift ($L = \frac{Confidence}{Support(rule\,head)}$;

$Support(rule\ head) = \frac{frequency(rule\ head)}{N}$) greater than 1 means rule body and head are not independent and rule body has an impact on the rule head.

2.2.4 Correlational and Comparative Analysis

Though CBA algorithm can extract unique patterns combining different behavioral data of different levels (e.g., high usage duration with the medium frequency of launch), it cannot present monotonic or linear statistical relation. To overcome the limitation, we did a correlation analysis. We used the nonparametric Spearman rank correlation (r_S) method as our data did not satisfy assumptions of the parametric test. In comparing the demographic data, we did a T-test (t) when data were normally distributed and in other cases, we did a nonparametric Mann-Whitney Test (U). As multiple comparisons can have false positive results, we adjusted p values using the false discovery rate approach.

2.2.5 Bayesian Network Analysis

Though correlation analysis can present the association between variables, it cannot reveal the direction of association between two variables. Therefore, to find the direction of the association, we built a Bayesian network [19] where each variable $V_i \in V$ is denoted by a node of a directed acyclic graph (DAG). In the development of the network, there are structural and parameter learning steps [19, 26] where the structural learning process is similar to the development of classical regression models [26]. For developing the network structure, we used the greedy hill-climbing algorithm which restarts randomly to avoid the local optima. Since the DAG of the network depends on several parameters such as the distribution of the data, we resample the Games app usage data and build the network 10,000 times to measure the strength of the associations and their direction. It can be noted that the strength of each arc was calculated keeping the rest of the network fixed and thus a relation between variables cannot be confounded by others.

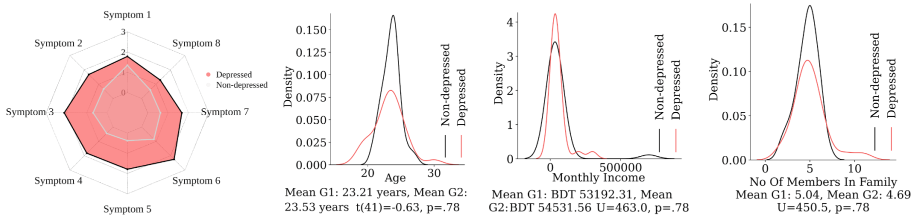
3 Results

3.1 Participants' Depression

Among the 60 Games app users, 46.67% (N = 28) were depressed and 53.33% (N = 32) were non-depressed (details about categorization process is in Sect. 2.2.1). Except for symptom 1 (Little interest or pleasure in doing things), every other symptom's mean score of the non-depressed group was below 1 (Fig. 2(a)) where score 1 presents the symptom's appearance not at all. Between these two groups, there was no statistically significant difference in age ($p = .78$) (Fig. 2(b)), monthly family income ($p = .78$) (Fig. 2(c)), and the number of family members ($p = .78$) (Fig. 2(d)) which could work as confounders in the relation between Games app usage and depression.

3.2 App Usage Behavioral Patterns' Association with Depression

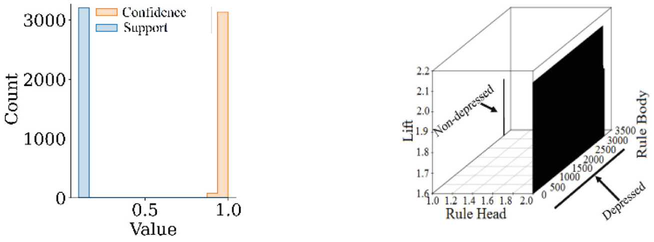
To extract patterns, we used all the behavioral data of weekdays, weekends, and 7 days. But to avoid combinatorial explosion, we set .1 as the minimum support and .9 as the



(a) Depressive symptoms' average score (b) Age (c) Monthly family income (d) Number of family members

Fig. 2. (a) Spider chart showing the difference between the depressed and non-depressed students regarding the mean score of each depressive symptom of the PHQ-8 scale where 0, 1, 2, and 3 scores present Not at all, Several days, More than half the days, and Nearly every day respectively. Kernel Density Estimation (KDE) plot showing the distribution of (b) age, (c) monthly family income (in BDT: Bangladeshi Taka), and (d) number of family members of the depressed (G1) and non-depressed (G2) students.

minimum confidence. In addition, as having a lift value of 1 presents that both the rule body and rule head are independent, we extracted only the rules which had a lift value higher than 1. Satisfying these thresholds, we found 3213 rules which were associated with depression. There were 21 rules where the rule head was non-depressed and in the remaining 3,192 rules, the rule head was depressed. Among these rules, the maximum support, confidence, and lift values were .17, 1.0, and 2.14 respectively (Fig. 3).



(a) Support and confidence (b) Lift

Fig. 3. Distribution of (a) support, confidence, and (b) lift values for the extracted 3,213 rules.

From the extracted behavioral patterns, we found that the higher duration per launching Games app was associated with depressive status (Table 1). For instance, the students whose weekdays' duration per Games app launch and 7 days' spending duration per app were high, they were more likely to be depressed. This behavioral pattern was observed among 17% of students and 91% of them were depressed students (Support = .17, Confidence = .91) (Rule 1, Table 1). The lift value (1.95) regarding this pattern was higher than 1 and this demonstrated that the rule body and rule heads were not independent. Also, we found when the students had medium entropy of using the apps for 7 days, duration per app was high on both weekdays and weekends, they were more likely to be depressed (Support = .17, Confidence = .91, Lift = 1.95) (Rule 5, Table 1).

Table 1. Top-5 (in terms of support and lift) Games app usage behavioral patterns of the depressed and non-depressed students. Con.: Confidence, Sup.: Support.

Depressed	Rules (Rule Body = > Rule Head)	Sup	Con	Lift
	1.{Duration_per_Launch,Weekday = High; Duration_per_App,7_days = High} = > {Depressed = Yes}	.17	.91	1.95
	2.{Duration_per_Launch,Weekday = High; Duration_per_Launch,7_days = High; Duration_per_App,7_days = High} = > {Depressed = Yes}	.17	.91	1.95
	3.{Duration_per_Launch,Weekday = High; Duration_per_App,Weekday = High; Duration_per_App,7_days = High} = > {Depressed = Yes}	.17	.91	1.95
	4.{Entropy,Weekend = Medium;No_of_Apps,7_days = Medium; Duration_per_Launch,Weekday = High} = > {Depressed = Yes}	.17	.91	1.95
	5.{Entropy,7_days = Medium; Duration_per_App,Weekday = High; Duration_per_App,Weekend = High} = > {Depressed = Yes}	.17	.91	1.95
Non-depressed	6.{Duration_per_Launch,7_days = Medium; Duration_per_App,Weekday = Medium} = > {Depressed = No}	.15	.90	1.69
	7.{Duration_per_App,Weekday = Medium; Launch_per_App,7_days = Medium} = > {Depressed = No}	.15	.90	1.69
	8.{Launch,Weekday = High; Launch,7_days = High; Duration_per_Launch,Weekday = Medium} = > {Depressed = No}	.15	.90	1.69
	9.{Launch,Weekday = High; Duration_per_Launch,Weekday = Medium} = > {Depressed = No}	.15	.90	1.69
	10.{Launch,7_days = High; Duration_per_Launch,Weekday = Medium} = > {Depressed = No}	.15	.90	1.69

On the other hand, the extracted rules regarding the non-depressed students presented that their spending duration per launch was medium (Rule 6 to 10, Table 1). For instance, the students whose spending duration per Games app launch was medium on 7 days and duration per Games app was medium on weekdays, were more likely to be non-depressed (Rule 6, Table1). This behavioral pattern was observed in the case of 15% of students (Support = .15) and 90% (Confidence = .90) of them were non-depressed. We also found that when the weekdays' duration per Games app and 7 days' frequency of launch per Games app became medium, they were also more likely (Confidence = .90, Lift = 1.69) to remain non-depressed (Rule 7, Table 1). Even when the participants had a high frequency of launching the apps, having a medium duration per launch presented a non-depressive status (Rule 10, Table 1).

3.3 Correlation Between Games App Usage and Depression

To explore the statistical relation, we did correlation analysis as discussed in Sect. 2.2.4. We found that the Games app usage data of weekdays and 7 days did not have any statistically significant relation with depression. In relation of depression with weekdays' spending duration ($r_s = .09$, $p = .48$), frequency of launching Games apps ($r_s = .08$, $p = .57$), and the number of used Games apps ($r_s = -.05$, $p = .71$) (Table 2), the p-value was much higher than the significance level .05. However, in case of weekends' spending duration ($r_s = .33$, $p = .026$), duration per launch ($r_s = .29$, $p = .044$), duration per app ($r_s = .39$, $p = .007$), frequency of launch per app ($r_s = .33$, $p = .026$), the p-value was less than .05 which demonstrated the significant association with depression score. This says that the students who spent higher time or have a higher frequency of launch per Games app on weekends were more likely to have a higher depression score.

Table 2. Relation of PHQ-8 score with usage data of Games apps. N denotes the number of users (who launched Games apps at least once in 7 days), in terms of a usage data. Coef.: Coefficient.

Usage data	Days	N	Coef. (r_s)	p
Duration	Weekdays	58	.09	.481
	Weekends	47	.33	.026
	7 days	60	.13	.339
Launch	Weekdays	58	.08	.573
	Weekends	47	.25	.089
	7 days	60	.07	.615
# of Apps	Weekdays	58	-.05	.713
	Weekends	47	-.14	.331
	7 days	60	-.15	.264
Duration per launch	Weekdays	58	.05	.721
	Weekends	47	.29	.044
	7 days	60	.11	.422
Duration per app	Weekdays	58	.12	.39
	Weekends	47	.39	.007
	7 days	60	.14	.290
Launch per app	Weekdays	58	.13	.326
	Weekends	47	.33	.026
	7 days	60	.12	.375
Entropy	Weekdays	31	-.31	.095
	Weekends	20	-.1	.667
	7 days	37	-.28	.098

3.4 Bayesian Network on Games App Usage and Depression

To understand the direction of the found significant associations (Table 2), we built a Bayesian network (Table 3 and Fig. 4) using the variables on weekends' data and by bootstrapping the data 10,000 times. We found that the probability (.61) of having an edge in the direction from Games app usage duration to PHQ-8 was higher than the direction from PHQ-8 to duration (.39) (Table 3). Similarly, the probability of having edges in the direction from duration per launch to PHQ-8 (.67) and from duration per app to PHQ-8 (.74) was higher than the probability in the reverse direction. This presents that the spending duration on Games apps, duration per Games app launch, and duration per Games app impacted depression (Fig. 4).

Table 3. Strength regardless of direction and strength in the specified direction. Gray-colored cells present arcs having more than 50% probability to appear in the specified direction.

Node (From)	Node (To)	Strength (Regardless direction)	Strength in the specified direction (From → To)	Node (From)	Node (To)	Strength (Regardless direction)	Strength in the specified direction (From → To)
PHQ-8	Duration	.13	.39	Duration	PHQ-8	.13	.61
PHQ-8	Entropy	.35	.60				
PHQ-8	Launch	.19	.69	Entropy	PHQ-8	.35	.40
PHQ-8	# of used apps	.44	.88	Launch	PHQ-8	.19	.31
PHQ-8	Duration per launch	.26	.33	# of used apps	PHQ-8	.44	.12
PHQ-8	Duration per app	.15	.26	Duration per launch	PHQ-8	.26	.67
PHQ-8	Launch per app	.10	.71	Duration per app	PHQ-8	.15	.74

Unlike duration, in the frequency of launching the Games apps, we found that the probability (.69) of having the edge in the direction from PHQ-8 to launch was higher than the probability of having the edge in the reverse direction (launch to PHQ-8: .31) (Table 3). In the same way, the probability of having the edge in the direction of PHQ-8 to the number of used apps (.88), launch per app (.71), and entropy (.60) were higher than the probability of having the edge in the reverse direction. This reveals that depression also impacted the frequency of launching of Games apps, number of used Games apps, launch per Games app, and entropy regarding Games app usage (Fig. 4).

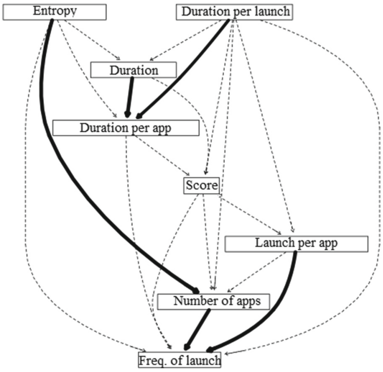


Fig. 4. Bayesian network presents the direction of the relation between PHQ-8 score and weekends' Games apps usage data.

From the Bayesian network (Fig. 4), it is also apparent that the frequency of launch node does not have any outgoing edge whereas the duration per app launch node has 6 outgoing edges which is the maximum among all nodes. This node is directly linked with the PHQ-8 score and also with the 5 behavioral markers, namely, duration, duration per app, number of used apps, launch per app, and frequency of launch. This presents that the duration per app launch can be a plausible target node to control the 5 behavioral markers and also depression.

4 Discussion

In our study on Games app users, we found a depression prevalence of 46.67% which is close to the depression prevalence of 47.3% found in a previous study conducted on Bangladeshi students [25]. Our analysis showing the negative impact of weekends' usual usage of Games apps on depression extends previous studies [24, 29] which used subjective data and found a negative impact of problematic gaming. To our best knowledge, this is the first study to present this impact using the objective app usage behavioral data and also using data of all the used Games apps by a participant. Our findings suggest raising awareness to reduce gaming time for their well-being, especially during this pandemic when gaming time increased significantly [11]. One of the plausible reasons for having a negative impact is that gaming makes a poor connection with family and friends [4] and this may have a significant negative impact on the students amid the pandemic. During alone time, people use smartphones to seek support [7] and on weekends, as students do not have classes and also as the pandemic restricted movement, higher usage of Games apps can present their effort to overcome loneliness through playing Games. Therefore, having a good connection with the parents, caregivers, or friends on the weekends may reduce their interest in the Games apps which in turn may reduce their spending time on Games. Moreover, our findings suggest parents and caregivers need to be aware of the weekend depression [20] since this has become a rising concern and also as we found a negative impact on weekends' Games app usage.

In our developed Bayesian network, we found some association in the opposite direction also, i.e., depression also impacts some weekends' behavior regarding Games apps. For instance, we found that depression has an impact on behavioral markers such as frequency of launching and number of used Games apps. In previous studies (e.g., [5, 24]), researchers showed how gaming is associated with depression. However, our analysis based on the Bayesian network where directed acyclic graphs were constructed 10000 times, showed the link in both directions depending on behavioral markers (e.g., duration, launch) which was unexplored even in subjective data-based studies [5, 24, 28]. Therefore, the insights from our findings can contribute to a better understanding of the smartphone usage behavior of the people [1, 15], especially the vulnerable group's Games playing behavior which can be potential in research to develop systems for better mental health. The plausible reasons for having such an impact of depression can be smartphone users' willingness to be distracted through apps upon facing negative emotions [21]. Higher launch and higher number of Games apps used by depressed students as shown in our study can present their multitasking behavior which can also present their distracting behavior. Therefore, these insights can facilitate in designing systems to regulate the Games app usage for promoting well-being.

Like the correlation analysis, in mined behavioral patterns through the CBA algorithm [16], we did not find link of depression with a single weekdays' or 7 days' behavioral marker. Instead, through mining rules, we found high usage of Games apps in terms of multiple behavioral markers was associated with depression. It is due to the fact that to find a relation with a class (e.g., depressed), the CBA algorithm [16] uses values of several variables in different combinations while the Spearman correlation [22] uses data of a single variable to find out the monotonic relation with another variable. This demonstrates the strength of the data mining technique in extracting Games app usage behavioral patterns of the students having psychological problems. In a study [23], researchers presented the application of mined behavioral patterns to develop machine learning models to identify depression with higher accuracy. However, Games app usage data was unexplored as features for the models. Our findings suggest that as Games app category's behavioral patterns are linked with depression, this app category's data can be leveraged to develop better computational models for real-time identification of depressed individuals.

From our developed Bayesian network, we found that the duration per Games app launch data has the maximum outgoing edges, and also this node is directly linked with depression. Hence, this can be a plausible target for limiting Games app usage behavior and also lowering depression. This finding extends the recent study on psychological problems where researchers discussed the interaction of depressive symptoms [3] presenting a plausible target for intervention. Moreover, as we found that the higher Games apps usage duration per launch had a negative relation with depression, Games developers may take this into account for the well-being of the students. They may integrate different interventions (e.g., intervention through the input [13]) which was found to be effective in minimizing app usage. But the Games should have the option to integrate students' self-defined rules since pre-defined intervention can create frustration [13].

5 Limitations

This study is limited by a small number of Games app users (N = 60) and 7 days' Games app usage data. In addition, due to having fewer participants in each sub-category of Games, we could not explore the sub-categories to analyze the relation with depression.

6 Conclusion

Using the objective app usage data, we explored the relation between depression and Games app usage behavior. From the Bayesian network-based analysis, we found that the relation is not in a single direction. Also, our mined class association rules through the CBA algorithm showed that depressed and non-depressed students have unique behavioral patterns. Insights from our findings can be potential for the caregivers to be aware of the negative impact of Games app usage. Researchers, developers, and healthcare professionals can also use these insights to design systems for well-being.

References

1. Ahmed, M.S., Ahmed, N.: Exploring unique app signature of the depressed and non-depressed through their fingerprints on apps. In: Lewy, H., Barkan, R. (eds.) PH 2021. LNICSSITE, vol. 431, pp. 218–239. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99194-4_15
2. Ahmed, M.: 86pc university students own smartphones in Bangladesh: Survey. https://en.prothomalo.com/youth/education/86pc-university-students-own-smartphones-in-bangladesh-survey. Accessed 21 Sept 2022
3. Briganti, G., Scutari, M., Linkowski, P.: Network structures of symptoms from the Zung Depression Scale. Psychol. Rep. **124**, 1897–1911 (2021)

4. Brigham Young University: Video games linked to poor relationships with friends, family (2009). https://www.sciencedaily.com/releases/2009/01/090123075000.htm
5. Brunborg, G.S., Mentzoni, R.A., Frøyland, L.R.: Is video gaming, or video game addiction, associated with depression, academic achievement, heavy episodic drinking, or conduct problems? J. Behav. Addict. **3**, 27–32 (2014). https://doi.org/10.1556/JBA.3.2014.002
6. Google Play most popular app categories 2021. https://www.statista.com/statistics/279286/google-play-android-app-categories. Accessed 23 Sept 2021
7. Diefenbach, S., Borrmann, K.: The Smartphone as a pacifier and its consequences: Young adults' smartphone usage in moments of solitude and correlations to self-reflection. In: Proceedings of the CHI 2019. ACM, New York, NY, USA (2019)
8. Drachen, A., Riley, J., Baskin, S., Klabjan, D.: Going out of business: auction house behavior in the massively multi-player online game glitch. Entertain. Comput. 5, 219–232 (2014)
9. UsageStatsManager. https://developer.android.com/reference/android/app/usage/UsageStatsManager.html. Accessed 21 Sept 2022
10. Hosen, I., Al-Mamun, F., Mamun, M.A.: Prevalence and risk factors of the symptoms of depression, anxiety, and stress during the COVID-19 pandemic in Bangladesh: a systematic review and meta-analysis. Glob. Ment. Health (Camb.) **8**, e47 (2021)
11. 75% of pandemic-driven increase in mobile gaming activity will persist indefinitely, according to new IDC and LoopMe report. https://www.idc.com/getdoc.jsp?containerId=prUS47906621. Accessed 21 Sept 2022
12. Ilhan, A.E., Sener, B., Hacihabiboglu, H.: Improving sleep-wake behaviors using mobile app gamification. Entertain. Comput. **40**, 100454 (2022)
13. Kim, J., Park, J., Lee, H., Ko, M., Lee, U.: LocknType: lockout task intervention for discouraging smartphone app use. In: Proceedings of the CHI 2019 (2019)
14. Kroenke, K., et al.: The PHQ-8 as a measure of current depression in the general population. J. Affect. Disord. **114**, 163–173 (2009)
15. Ahmed, M.S., Rony, R.J., Hasan, T., Ahmed, N.: Smartphone usage behavior between depressed and non-depressed students: an exploratory study in the context of Bangladesh. In: Adjunct Proceedings of the UbiComp-ISWC 2020 (2020)
16. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In: Proceedings of the KDD 1998, pp. 80–86. AAAI Press (1998)
17. App usage - manage/track usage. https://play.google.com/store/apps/details?id=com.a0soft.gphone.uninstaller. Accessed 21 Sept 2022
18. Mok, L., Anderson, A.: The complementary nature of perceived and actual time spent online in measuring digital well-being. Proc. ACM Hum. Comput. Interact. **5**, 1–27 (2021)
19. Nagarajan, R., Scutari, M., Lèbre, S.: Bayesian Networks in R (2013)
20. Owl, B.: Weekend depression. https://www.kindmindonline.com.au/blog/weekend-depression-ecmh9. Accessed 21 Sept 2022
21. Sarsenbayeva, Z., et al.: Does smartphone use drive our emotions or vice versa? A causal analysis. In: Proceedings of the CHI 2020 (2020)
22. Chalmer, B.J.: Understanding Statistics. CRC Press, Boca Raton, FL (1986)
23. Xu, X., et al.: Leveraging routine behavior and contextually-filtered features for depression detection among college students. Proc. ACM IMWUT **3**, 1–33 (2019)
24. Yazici, Z.N., Kumcagiz, H.: The relationship between problematic online game usage, depression, and life satisfaction among university students. Educ. Proc. Int. J. **10**, 27–45 (2021)
25. Koly, K.N., et al.: Prevalence of depression and its correlates among public university students in Bangladesh. J. Affect. Disord. **282**, 689–694 (2021)
26. Scutari, M., Auconi, P., Caldarelli, G., Franchi, L.: Bayesian networks analysis of malocclusion data. Sci. Rep. **7**, 15236 (2017). https://doi.org/10.1038/s41598-017-15293-w

27. Hunt, M.G., Marx, R., Lipson, C., Young, J.: No more FOMO: Limiting social media decreases loneliness and depression. J. Soc. Clin. Psychol. **37**, 751–768 (2018)
28. Wei, H.-T., Chen, M.-H., Huang, P.-C., Bai, Y.-M.: The association between online gaming, social phobia, and depression: an internet survey. BMC Psychiatry **12**, 92 (2012)
29. Lin, H.-C., Yen, J.-Y., Lin, P.-C., Ko, C.-H.: The frustration intolerance of internet gaming disorder and its association with severity and depression. Kaohsiung J. Med. Sci. **37**, 903–909 (2021)
30. McMahan, T., Parberry, I., Parsons, T.D.: Modality specific assessment of video game player's experience using the Emotiv. Entertain. Comput. **7**, 1–6 (2015)
31. Caplan, S., Williams, D., Yee, N.: Problematic Internet use and psychosocial well-being among MMO players. Comput. Hum. Behav. **25**, 1312–1319 (2009)
32. Kahn, A.S., Ratan, R., Williams, D.: Why we distort in self-report: Predictors of self-report errors in video game play. J. Comput. Mediat. Commun. **19**, 1010–1023 (2014)
33. Opoku Asare, K., et al.: Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: Exploratory study. JMIR mHealth uHealth **9**, e26540 (2021)

mHealth for Medication and Side Effect Monitoring: Patients' Attitudes Toward Smart Devices for Managing Oral Chemotherapy During Lung Cancer Treatment

Anna N. Baglione[1](✉), Sarah Tolman[1], Chloe Dapaah[1], Danielle Johnson[1], Kristen J. Wells[2], Richard D. Hall[1], Ryan D. Gentzler[1], and Laura E. Barnes[1]

[1] University of Virginia, Charlottesville, VA, USA
{ab5bt,sat2ew,cdd9zru,dbj4njw,lb3dp}@virginia.edu, {rdh3q, rg2uc}@uvahealth.org
[2] San Diego State University, San Diego, CA, USA
kwells@sdsu.edu

Abstract. In recent years, new treatments have become available which have improved survival rates in lung cancer patients. One promising treatment option is the rapidly growing field of oral targeted therapies, which employs drugs that interfere with specific molecules involved in the growth, progression, and spread of cancer. However, these therapies can cause a variety of symptoms and adverse events that can impair quality of life. mHealth technologies may help individuals with lung cancer better track their side effects and manage medications on a day-to-day basis. However, understanding patients' attitudes toward smart devices such as smartphones, smartwatches, and smart pill bottles, as well as their specific needs when using these devices, is critical before design and deployment studies of medication adherence can be carried out. In this study, we conducted interviews with 9 individuals with stage III-IV lung cancer at a National Cancer Institute-designated comprehensive cancer center in the Mid-Atlantic region of the United States to assess the feasibility of using such devices for managing medication and medication related side-effects. We evaluated patients' attitudes towards the design and function of smart devices and how these devices fit into their daily life. Our results may help clinicians and researchers to co-develop effective mHealth system deployments for side effect and medication management in oncology populations.

Keywords: mHealth · smartphone · medication adherence · side effect · cancer

1 Introduction

Lung cancer is the second most common type of cancer and the leading cause of cancer death worldwide, with over 2 million cases newly diagnosed each year [14]. The significant impact of lung cancer on the global population has led to the development of

---

C. Dapaah and D. Johnson—Authors contributed equally.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 570–583, 2023.
https://doi.org/10.1007/978-3-031-34586-9_38

new targeted oral anticancer medications, which patients tend to prefer for their convenience over intravenous chemotherapy [6, 26]. While promising for survival outcomes, these new therapies are commonly associated with adverse events (AEs) such as rashes or edema. AEs can lead to worsening symptoms, dose reductions, and even medication discontinuation if left undetected or untreated [4]. Researchers and clinicians are increasingly seeking more accurate ways to track medication-taking, monitor side effects, and detect possible AEs among patients taking oral anti-cancer medications at home, such as individuals with lung cancer.

Devices, such as smartphones, wearable sensors (e.g., smartwatches) and medication event monitoring systems (MEMS), enable direct, unobtrusive collection of clinically relevant behaviors in-situ. Mobile health (mHealth) and human-computer interaction (HCI) studies have shown that these "smart" devices are less prone to errors than traditional self-reports [2] and have established the usefulness of smart devices for medication and symptom tracking in daily life [3, 19]. At the intersection of HCI, mhealth, and oncology, smart devices have been shown to encourage medication adherence to oral chemotherapy [15], help patients feel more in control and informed about their care [11, 20, 21], and help clinicians feel better able to monitor patients' symptoms and tailor treatment accordingly [21]. However, the use of smart technologies for medication and symptom tracking in practice is not without its challenges. A 2017 study of medication adherence technologies such as smartphone apps among older adults showed that adherence was impacted both by participants' schedules and the symptoms they experienced [19]. Further, a study of medication tracking among patients with atrial fibrillation uncovered issues such as the inability of smart pill bottles to integrate into patients' existing routines [25]. These challenges highlight the importance of further investigation into patients' perceptions of smart device use during treatment.

Despite the promise of smart devices for symptom and medication management, the majority of studies in mHealth and oncology have focused on physical activity tracking for breast cancer patients [13, 16]. Comparatively few studies to date have focused medication and symptom tracking. Perhaps even fewer have focused exclusively on lung cancer patients, with a handful of notable exceptions. LuCApp is a mobile application for patients with lung cancer that sends automated reminders to complete symptom logs as well as questionnaires related to quality of life and support needs [5]. A proposed randomized controlled trial for the app will examine the impact of side-effect tracking on quality of life. A randomized controlled trial has also been proposed for SYMPRO-Lung, a web application for lung cancer patients that leverages patient-reported outcomes (PROs) for symptom monitoring [1]. We draw inspiration from these and other works as we address the following research question **(RQ):** ***What attitudes do patients with lung cancer have toward smart device use for managing their medications and tracking their symptoms?*** In this paper, we present the results of a cross-sectional, qualitative study in which we conducted semi-structured usability interviews with 9 individuals with stage III-IV lung cancer receiving treatment at a large university cancer center in the Mid-Atlantic region of the United States. Our results give insight into patients' preferences and priorities regarding the use of smart devices as part of their self-management routines during cancer treatment.

2 Methods

This study was approved by the Institutional Review Board for Health Sciences Research (IRB-HSR) at our university, and the study was conducted in accordance with the Declaration of Helsinki and Good Clinical Practice standards. Patients provided written informed consent prior to enrollment and participation.

2.1 Recruitment

Using purposive sampling, we recruited patients at a National Cancer Institute (NCI)-designated comprehensive cancer center in the Southeastern United States. All patients were 18 years of age or older and were being treated for advanced stage non-small cell lung cancer (NSCLC) with *EGFR* mutations or *ALK* gene rearrangements and were receiving oral targeted therapies (tyrosine kinase inhibitors [TKIs]) as part of their treatment. Patients were first identified for inclusion by the sixth and seventh authors, who are practicing oncologists. The first author attempted to contact prospective participants both in clinic and via telephone calls, and provided interested individuals with a secure, electronic consent form to sign. We approached 40 patients in total, 23 of whom either explicitly declined prescreening or were unreachable after one or more attempts to contact them. 17 patients agreed to prescreening, and 11 ultimately consented to participate in the study. Two participants did not respond to study coordinators' efforts to schedule the study interviews after consenting, bringing the final number of participants to 9.

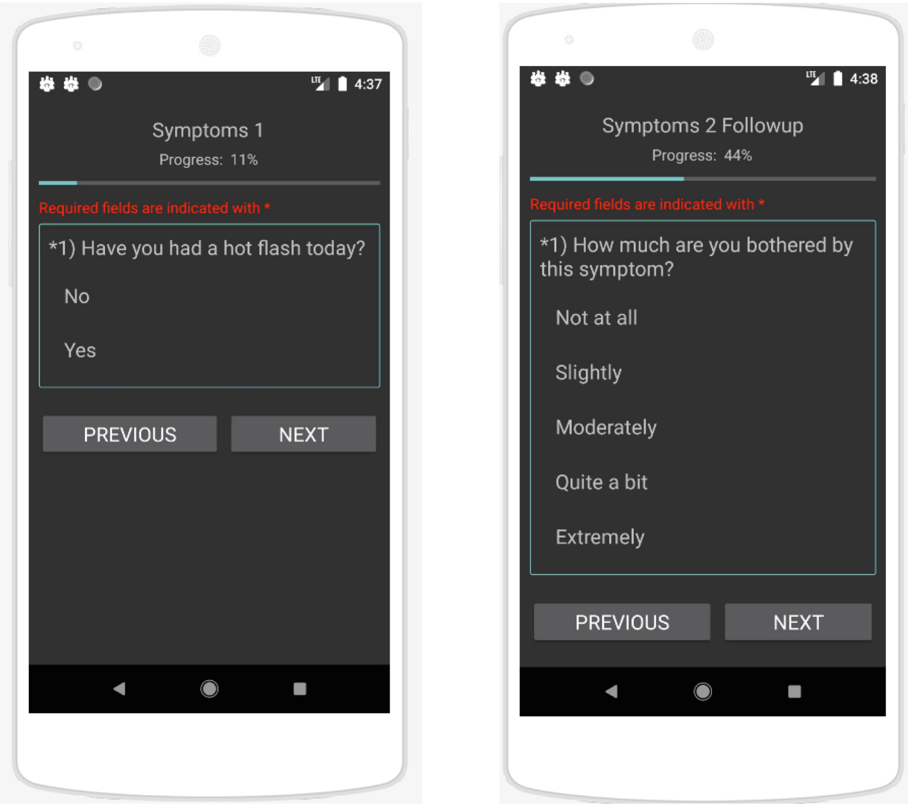
2.2 Data Collection

Using a standardized interview guide, we conducted semi-structured interviews with 9 participants between September 2020 and July 2021. Out of an abundance of caution during the COVID-19 pandemic, interviews were conducted by one interviewer remotely via a HIPPA-compliant version of Webex[1] using a standardized semi-structured interview guide. We administered a secure, online demographics survey via Qualtrics at the end of each interview. The first interview (with P1) focused on smartphone use and an interactive demonstration (demo) of a smartphone app emulator. P1's interview informed the design of subsequent interviews, which included an interactive demo of a smartwatch app emulator and a researcher-guided demo of a smart pill bottle cap in addition to the original smartphone app demo. In this section, we describe our process for each device demo in detail.

Smartphone. We created a high-fidelity prototype of Sensus [27], a smartphone application that gathers passively sensed indicators of human health and behavior (e.g., location, heart rate, and skin temperature). In deployment studies, Sensus can also be used to gather real-time participant feedback using *ecological momentary assessment (EMA)*, a method of gathering data in which participants are polled in real time in order to avoid recall bias [22]. EMAs are commonly delivered via digital methods such as text messages or push notifications from mobile apps. We created a Sensus study protype with

---

[1] WebEx; https://www.webex.com/.



Symptoms 1

Progress: 11%

Required fields are indicated with *

*1) Have you had a hot flash today?

No

Yes

Symptoms 2 Followup

Progress: 44%

Required fields are indicated with *

*1) How much are you bothered by this symptom?

Not at all

Slightly

Moderately

Quite a bit

Extremely

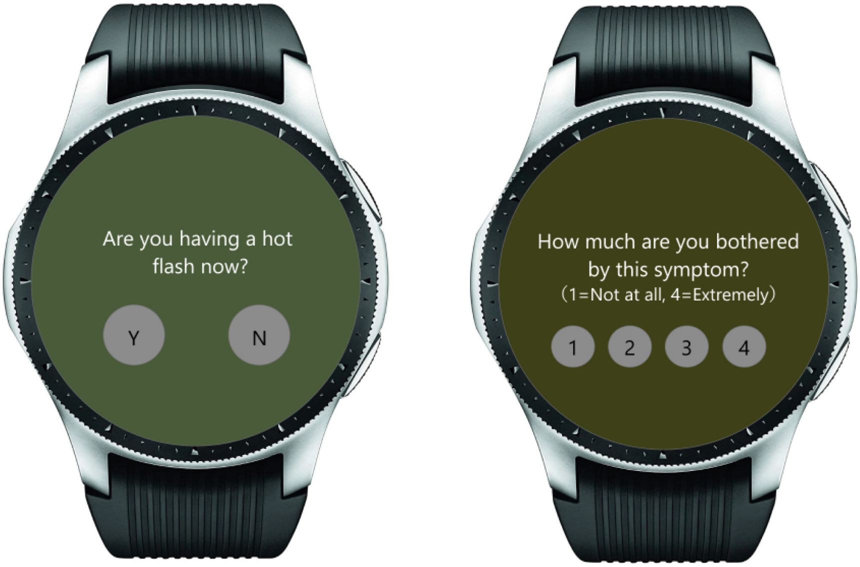
Fig. 1. Sensus prototype showing EMA survey about symptoms.

EMA surveys about participants' quality of life (e.g., sleep, symptoms, and side effects) and day-to-day activities (e.g., location and socialization). We then loaded the Sensus protocol via Appetize.io[2], an app demo platform for the web; the prototype is shown in Fig. 1. During the interview, we used screen sharing to show participants how to scroll and select questions within the prototype; this was necessary, as using a mouse for these tasks is very different from swiping and tapping with one's finger on a real smartphone. We then asked participants to use the prototype via screen sharing to practice answering the survey questions. Finally, we asked participants 15 follow-up questions that covered their perceptions about the experience of filling out surveys on a smartphone, the relevance of the smartphone to their current medication management routine, and their willingness to use Sensus on a smartphone for a long period of time.

Smartwatch. We used a web-based prototype of a smartwatch application (shown in Fig. 2) that was preloaded with quality-of-life questions, as well as questions regarding activities (e.g., "Are you exercising right now?"). We explained to participants how the smartphone could be used to send EMAs and notifications to the smartwatch. We then

---

[2] Appetize.IO; https://appetize.io/.



Are you having a hot flash now?

How much are you bothered by this symptom?
(1=Not at all, 4=Extremely)

Fig. 2. Smartwatch app prototype showing EMA survey about symptoms.

asked participants to use the smartwatch app prototype via screen sharing to practice filling out EMAs. Finally, we asked participants 15 follow-up questions that covered their perceptions about the experience of filling out EMAs on a smartwatch, the relevance of the smartwatch to their current medication-taking routine, and their willingness to use the app on a smartwatch for a long period of time.

Smart MEMS Cap. For the final segment of the interview, we used screen sharing to demonstrate the use of the RxCap[3], a bluetooth-enabled MEMS cap that records each time the cap is unscrewed as a medication-taking event. Since most participants had not seen or used a smart pill bottle before, we first explained the purpose and function of the cap. We then showed participants how one would remove the cap to take medication, and how the cap would blink and beep upon removal. We also explained how the cap could be connected to an application on the user's smartphone to help keep track of when they took their medication. In order to gain a better understanding of participants' medication-related needs, we asked questions about the types of medications participants were currently taking, the frequency with which they took them, their preferred storage method (e.g., pill box; original bottle) and location (e.g., in the bathroom; on a nightstand), and what kinds of alerts or reminders they used to help them remember to take their medications (e.g., app; phone alarm) We then asked 15 follow-up questions about the relevance of the cap to their current medication management routine and their willingness to use a smart pill bottle to store their medications for a long period of time.

---

[3] RxCap; https://rxcap.com/.

2.3 Data Analysis

Interviews lasted between 30 min and 1.5 h and were audio recorded. The first and second author transcribed the interviews verbatim. The first author then applied initial codes from the interview transcripts to develop a preliminary code-book. She then worked with the second, third, and fourth authors use an iterative, inductive approach to refine the initial codes, develop new codes, and extract the overall themes. Iterations continued until all coders reached consensus.

Demographic data were gathered and analyzed by the first and second author, who extracted the data from the secure Qualtrics survey and ran descriptive statistics (e.g., means, standard deviation, frequencies) using Microsoft Excel.

3 Results

3.1 Participant Demographics

Participants' ages ranged from 33 to 86 years ($\mu = 60.2$ years; $\sigma = 15.8$ years), with a gender distribution of 6:3 (female: male). Most participants self-identified as White (8/9; 88.9%), with one participant indicating they were Asian (1/9; 11.1%). Among those who reported their ethnicity (8/9; 88.9%), one participant reported being of Hispanic ethnicity (1/8, 12.5%). All participants (9/9; 100%) had been diagnosed with lung cancer at least 6 months prior to the study. Among those who reported their lung cancer stage (8/9; 88.9%), the majority were diagnosed with stage IV (7/8; 87.5%), followed by stage III (1/8; 12.5%). One-third were former smokers (3/9; 33%). Among the former smokers, the mean number of years of tobacco use was $\mu = 21.7$ years ($\sigma = 2.4$ years).

3.2 Interview Study Findings

Participants' attitudes, concerns, and needs regarding smart device use spanned four key thematic areas: *device and application design, lifestyle*, and *abilities, and obligations.* In this section, we delve into each of these themes in detail to answer our original research question **(RQ):** *What attitudes do patients with lung cancer have toward smart device use for managing their medications and tracking their symptoms?* In this section, we present participants' attitudes and concerns toward the individual devices as well as their needs and preferences for notifications they might receive from any device. We also describe how participants' personal and social obligations affect their willingness to use smart devices. Finally, we highlight the roles of self-efficacy and obligation to self and others in motivating smart device use.

Device and Application Design. *Smartphone and Smartwatch.* Participants appreciated the smartphone and smartwatch for their compact size and ease of use. Both P1 and P11, for instance, expressed a preference for the smartphone over bulkier technologies. In P11's words, *"It doesn't force me to have to go to the computer."* While the smartphone was familiar to most participants, the smartwatch was not, and participant opinions on the smartwatch were divided. Some were drawn to the smartwatch because they could input data directly on their wrist in a discreet way. P4, for instance, described how his personal smartwatch was useful for discreetly checking messages while at work. P11, however, worried that smartwatch notifications in particular were a privacy risk. Given that the watch must be worn at all times rather than kept aside in a purse or pocket, the watch has the potential to draw unwanted attention to private messages in social settings: *"Notifications will bother me more on my watch than on my phone…I'm not sure if I would like going out for dinner, and all of a sudden noticing that my wrist is lighting up with a message and somebody across the table says, oh, you have a message in your wrist… [Or] what if I'm having an important meeting with somebody?"* P4 also found the smartwatch somewhat intrusive (despite regularly using his own), and preferred to keep the device in the background as much as possible: *"It requires an answer right then and there. Personally I don't like inputting [data]. I see [the watch] as more of a way to receive information."*

Participants also mentioned several design-related needs and concerns, with regard to the smartwatch and smartphone. Some worried the surveys were too long and would become cumbersome, or that the device's battery might drain too quickly due to running an app. Attitudes about device size were divided; one participant found the smartwatch screen too small and hard to navigate, while another, P6, found the watch to be too big for regular use: *"I'm not a big fan of wearing much on my wrist..I would be inclined to forget wearing it, I'm afraid, because it's bulky… I do not like intrusive technology."* Several participants wanted changes to the surveys, including more aesthetically-pleasing color schemes and the ability to comment on the frequency of their symptoms. Participants also wanted to complete surveys at their own time and pace, with several wanting to set their own notification schedule.

Pill Bottle. Of the three smart devices we studied, the smart pill bottle received the least support from participants. Most participants were taking multiple prescription medications in addition to multiple vitamins and supplements, and disliked the smart pill bottle because it could only hold one type of pill at a time. As P11 described, they tended to prefer divided pill boxes for everyday use: *"I think the box that has the separated days would be more useful because it would be recording not only that you took it, but [when] you took it…So the data that you would record would be more complete."*

Notifications. In general, participants valued their privacy and peace. They wanted notifications to be unobtrusive and discreet, especially in public settings such as the workplace. Participants' preferences, in this regard, were very personal. Some participants preferred to leave most notifications off and found them "annoying". P8 was willing to receive vibrations only, in keeping with his work obligations: *"A vibration is best, because a lot of times I'm in management meetings. Obviously, we all have our phones turned down."* Others, such as P7, were willing to receive audible "dings" on any device, provided they were not overly loud or repetitive: *"I would want it to be quieter and more subtle, …and also not persistent. So one notification is fine, [but] five notifications would not be fine… I would not want to have to keep seeing it. I'm [also] notorious for clearing out my notifications, and in fact I turn off a lot of my notifications because it's a privacy issue to me."* P6 expressed a similar preference: *"I set my alarms for my meds so I would definitely [want notifications]…I would probably have it be a single ding…I don't want anything irritating like 'DA, DA, DA, DA!' Just like a single ding would work."*

Several participants also expressed a desire for survey notifications to be integrated with their electronic health record apps, so that all their health-related notifications showed up in one place. For example, P4 described how he would be more likely to take surveys when checking for appointments in MyChart[4], a popular electronic health record (EHR) application. Importantly, participants wanted to receive notifications only when absolutely needed, given the burden of time their cancer treatment already placed on them. For instance, P11 described how her view of time had changed since her diagnosis and re-emphasized that notifications should be minimally disruptive, especially during family and personal time: *"When you have cancer, too many things seem too trivial, and you want to concentrate every day on using your time in the best possible manner. So it's funny, I'm now quite bothered by all these notifications that come to me about celebrities…but I do want to get notifications if my sons do something. So I think it depends. I would say not too many; enough notifications that we can do this [study], but not unnecessary notifications."*

Lifestyle. Participants' lifestyles heavily influenced their attitudes towards smart device use. P6, echoing many other participants, cited her familiarity and current use of smartphones as a reason they would be willing to use the Sensus app as part of a future study: "Like a lot of people, I use my phone more than any other device." Participants were less familiar with the smartwatches. Even P4, who owned a smartwatch, was concerned about learning to use a different type of smartwatch when he already owned one that worked well for him: "I wouldn't like it. I prefer my watch and the features my watch has, I've already gotten used to it. I wouldn't want to learn a whole new system, and I'm assuming if it's a research watch I wouldn't be able to install any of my own apps on the watch anyways."

Participants' schedules and levels of flexibility varied, though most concluded that they were more available on weekdays than on weekends. Weekends were often reserved for family time and other social activities (e.g., hosting friends or attending church services). For instance, P1 described the importance of time with her husband: *"We're out hiking or doing projects. I'm less likely to think about doing something like a survey."* Similarly, P6 did not want to be interrupted during her valued social time: *"Generally we are busier on the weekends with catching up with friends and family…I set an alarm on my phone for taking my medication, and then honestly if we're out socializing, I end up snoozing the alarm and snoozing the alarm… I take it within an hour or two. But… [at] nine o'clock on Saturday night I don't generally want to be interrupted with a reminder about something, and I think I'll feel the same way about the surveys."* Several participants also mentioned that their activities could put them out of range for receiving push notifications from a study device (e.g., alarms or reminders to take a medication dose). For instance, before the COVID-19 pandemic forced many people to stay at home, P8 liked to go hiking on the weekends in remote areas with little-to-no cellular service: *"Right now we're all hunkered down [during the pandemic]. I used to be out of range of any technology if I was skiing or backpacking. I could be gone for 13 days, out of range."*

---

[4] MyChart; https://www.mychart.com/.

Throughout the interviews, participants repeated their commitment to habits and routines as a major influencing factor on their attitudes toward smart devices. Several participants, such as P6, reported that using the devices would *"just become a part of a daily routine"*. P9 even likened smart device use to taking medicine regularly*: "What's the difference of using the app every day and then using the medicine every day? I don't know if there would be a difference."* While participants felt the smartwatch and smartphone could fit into their existing routines, they did not feel the same way about the pill bottle. The reasons participants gave for not wanting to use the pill bottle were as much of a lifestyle concern as a design concern. Namely, managing multiple pills had driven participants to establish longstanding, personalized medication-management routines that already worked well for them. P2 put it simply: *"I don't need [the smart pill bottle] – "I have no problem with what I'm doing now."*

Abilities. Participants exhibited varying degrees of self-efficacy with regard to using smart devices as part of their medication management routine. P2 noted that, while she was confident, she could use the real, physical devices, the screen size and difficult scrolling in the online prototype were challenging for her. Others, such as P1, expressed confidence in their own technical skills, but were doubtful that other people, especially in older age groups, would be able to use the devices: *"I don't think I would have any problem at all using it. I could see if it was my mother, I would have to train her how to use it. But the way it's set up, for anyone that uses apps, it's pretty obvious what to do."* Still others believed they would need significant support from the study team or another support person. For instance, P7 asked, "*Are you going to train me really well in how to use that smartwatch?"* P3 believed she could use the smartphone if she received outside help from her grandchildren or a tutor: *"I might even hire a tutor to help me…I'm not that great on a smartphone for sure."*

Obligations. Besides force of habit, participants cited obligations to their doctors and to themselves as a factor influencing their willingness to use devices. Participants took great pride and responsibility in managing their health as best they could during treatment. They valued smart devices for their ability to track symptoms over time and wanted the ability to share this information with their providers. For instance, P1 described how an app could help her keep an accurate record of her symptoms and side effects to present to her oncologist: *"I'm not going to call my doctor and say, oh I had a mouth sore today. But if I had an app, if I was supposed to use it every day, I would put it into the app."* Likewise, P7 would be willing to fill out surveys more frequently if it helped with symptom management: *"If [there were] questions that I felt like were important for me and my doctor to know, I would do it five times a day… I would do it… if I thought it would help manage my symptoms."*

4 Discussion

Our study highlights a number of considerations and challenges for designers at the intersection of mHealth, medication adherence, and oncology. Participants' openness to using certain devices may be mitigated by their familiarity with the device, confidence in their technical abilities, work and social obligations, and even their fashion preferences. Moreover, fundamental challenges, such as whether participants will be available and within range for receiving push notifications, will necessarily influence the design and deployment of medication management technologies. Based on these findings, we present several practical design considerations in the following sections.

4.1 Give participants agency over their notifications

Notifications and reminders to take one's medication play an important role in interventions for individuals with cancer. Prior work has demonstrated the feasibility of using information from smart devices to inform the delivery of missed dose messages via EHRs, such as MyChart [25]. This approach is a promising step towards more personalized notifications and interventions. Prior research with cancer patients taking oral chemotherapy has emphasized the importance of taking the user's schedule into account when delivering reminders [23]. In a similar vein, participants in our study overwhelmingly expressed a desire to control the timing and format of notifications as a condition of using the smart devices in their regular medication-taking routine. We recommend that designers of smart device applications for medication management give participants a range of options for customizing their notification frequency from within the app. Designers might consider setting a default schedule based on times of day most commonly associated with medication taking and alertness, then enabling participants to customize this schedule as needed. For instance, 8 AM and 8 PM often correspond with morning mealtimes and evening bedtime routines, respectively. Ideally, users would be presented with a screen that allows them to set the exact day(s) and time(s) they would like to receive notifications, as well as the type of notification (e.g., vibration, beep, or banner) for each day and time. Giving participants agency over their notifications in this manner is a small cost for designers, but a major step toward protecting participant privacy and ensuring notifications are well-integrated into participants' daily routines (rather than intruding on them).

4.2 Advocate for Better Avenues for Secure Sharing of Patient-Entered Data with Clinical Care Providers

Prior works have established patients' openness to sharing information such treatment satisfaction and adverse effects with their clinicians via apps, provided that the information can be used to complement their treatment [12]. Our participants shared this attitude. Given the significant physical and emotional toll of their lung cancer treatment, participants had a vested interest in being able to review and share as much of their passively- and actively-sensed health data as possible with their doctors. Researchers have called for better infrastructure for secure information sharing between patients and providers in the context of mHealth for cancer care, given the limitations of current modalities [18]. Yet, this remains an open challenge. Electronic Health Records (EHRs) are the gold standard for health information and records management in our digital world, yet they are primarily designed for displaying information entered by clinical care providers (e.g., patient lab values and test results) and for facilitating basic secure messaging between patients and providers. EHRs in their current form are not equipped to receive and process information from consumer devices, such as smartwatches, or from custom medication and symptom tracking apps, in part due to strict requirements imposed by laws such as HIPAA [9] and the HITECH Act [8]. Moreover, building a standalone application that securely transmits patient-entered data to the patient's healthcare provider via an existing EHR platform would be an enormous challenge. Such an application would not only need to comply with current market standards for secure health information sharing, such as Health Level 7 (HL7) [10], but would require direct collaboration with leading EHR vendors. Additionally, such an application would need to navigate the gaps left by HIPAA with regard to digital healthcare tools [24].

Designers seeking a short-term solution for patients who wish to share information with their doctors could provide in-app visualizations at different timescales, such as daily charts of the times of day a patient took their medication, or weekly and monthly graphs of adherence percentages over time. These should be easily-exportable to images that participants could save to their smartphone and share with their doctors manually in-clinic during routine appointments. Designers taking this approach should consult with both patients and clinicians when designing such visualizations, to ensure they are clear and concise for both parties. As a long-term solution, designers should consider advocating for improved guidance on patient-to-clinician information sharing via digital technologies, at the national level, and should seek out long-term collaborations with EHR vendors and smart device manufacturers where possible.

4.3 Ensure Participants have Access to Adequate Support Resources During Deployment Studies

Prior work in mHealth has shown that patients with lung cancer may feel they are lacking sufficient support and self-management skills, in regards to their disease [17]. Several participants echoed these concerns specifically with regard to using smart devices. Indeed, many expressed a hesitancy to use smart devices due to their perceived lack of technical proficiency. Participants also expressed a need for extensive support from the study team, should they choose to use the smart devices in a future deployment study. To increase participants' confidence, we suggest providing a comprehensive technology use manual and other written educational materials that describe how to use each study device. We also recommend that study team members review these materials with participants, and provide in-depth demonstrations of each device. We note that technological concerns are likely to arise during deployment. To adequately address these concerns, we also recommend providing the participant with a specific study contact designated to addressing and supporting individual technology needs.

5 Limitations

This study is novel given its focus in evaluating perceptions of smart technology among individuals living with advanced lung cancer. However, it does have limitations, including the small sample size. We faced several recruitment challenges during this study. Cold-calling potential participants was largely unsuccessful. Among those who responded to cold calls or were willing to speak to us in clinic, most declined. Their reasons included disease burden (e.g., fatigue from treatment), busyness due to participating in other research studies, and lacking a computer for the study interview. Whether these challenges were unique to our study population is outside the scope of this paper; however, we recommend that future studies cast a broad recruitment net across multiple treatment facilities if possible.

Like most studies conducted during the COVID-19 pandemic, we also faced challenges in adapting our study activities to be fully remote. While we were able to conduct recruitment in clinic by taking many precautions such as masking, we opted to conduct the interviews remotely to reduce the risk of transmission of COVID-19 to participants. This decision fundamentally altered who we enrolled in the study. Our remotely conducted interviews required a personal computer (PC) with a mouse and microphone, preventing those without a PC from participating. Moreover, those who did participate did not get the in-person experience of physically interacting with the study devices. Additionally, we struggled to recruit a racially diverse sample. We also urge researchers in the field to continue to increase efforts to recruit diverse participants. Given the significant racial and social disparities present in the incidence of lung cancer [7] and other diseases, diverse samples are necessary for designing mHealth tools that serve as many patients as possible.

6 Conclusion

In this study, we presented finding from interviews with 9 individuals with lung cancer on patients' attitudes and needs towards mHealth devices for side effect and medication management. Our findings shows that patients' motivations for using these devices are dependent on a number of important factors, including the devices' design, patients' lifestyles and existing habits, their unique abilities and feelings of self-efficacy, and their preexisting obligations. These findings may help clinicians and researchers to co-develop effective deployments of mHealth systems for side effect and medication management in oncology populations. Specifically, our results will help study developers to identify which features patients find most valuable for specific "smart" devices, such as the ability to view and download one's data from a mobile app. More work is needed to see if our results and implications for design overlap with findings for other oncology populations.

Acknowledgements. This research was supported by the following sources: the National Cancer Institute Cancer Center Support Grant 2P30CA044579–31 (to A. N. Baglione, R. Hall, and R. Gentzler); the National Cancer Institute, under award number R01CA239246 (to Dr. L. E. Barnes, Dr. J. Gong, and Dr. K. J. Wells); and the Lung Cancer Research Foundation (LCRF), supported by Pfizer (to Dr. R. Gentzler). The content in this paper is solely the responsibility of the authors and does not necessarily represent the official views of the aforementioned sources.

We thank Cristian Garcia Alcaraz and Manuel Gonzales for contributing to the development of the interview questions and demographic surveys. We also thank Xishi (Ethan) Zhu and Xue (Shirley) Wu for their work on the smartwatch mobile app, and Dr. Jackey Gong for his supervision of the app's development.

References

1. Billingy, N.E., et al.: SYMptom monitoring with Patient-Reported Outcomes using a web application among patients with Lung cancer in the Netherlands (SYMPRO-Lung): study protocol for a stepped-wedge randomised controlled trial. BMJ Open **11**, e052494 (2021). https://doi.org/10.1136/bmjopen-2021-052494
2. Burney, K.D., Krishnan, K., Ruffin, M.T., Zhang, D., Brenner, D.E.: Adherence to single daily dose of aspirin in a chemoprevention trial: an evaluation of self-report and microelectronic monitoring. Arch. Fam. Med. **5**, 297 (1996)
3. Caldeira, C., Bhowmick, P., Komarlingam, P., Siek, K.A.: A state-based medication routine framework. In: CHI Conference on Human Factors in Computing Systems, pp. 1–16. ACM, New Orleans LA USA (2022). https://doi.org/10.1145/3491102.3517519
4. Califano, R., et al.: Expert consensus on the management of adverse events from EGFR tyrosine kinase inhibitors in the UK. Drugs **75**(12), 1335–1348 (2015). https://doi.org/10.1007/s40265-015-0434-6
5. Ciani, O., et al.: Lung Cancer App (LuCApp) study protocol: a randomised controlled trial to evaluate a mobile supportive care app for patients with metastatic lung cancer. BMJ Open **9**, e025483 (2019). https://doi.org/10.1136/bmjopen-2018-025483
6. Geynisman, D.M., Wickersham, K.E.: Adherence to targeted oral anticancer medications. Discov. Med. **15**, 231–241 (2013)
7. Harper, S., Lynch, J., Meersman, S.C., Breen, N., Davis, W.W., Reichman, M.E.: An overview of methods for monitoring social disparities in cancer with an example using trends in lung cancer incidence by area-socioeconomic position and race-ethnicity, 1992–2004. Am. J. Epidemiol. **167**, 889–899 (2008). https://doi.org/10.1093/aje/kwn016
8. Health Information Technology for Economic and Clinical Health (HITECH) Act, Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009 (ARRA), Pub. L. No. 111–5, 123 Stat. 226 (Feb. 17, 2009). (full-text), codified at 42 U.S.C. §§300jj et seq.; §§17901 et seq
9. Health Insurance Portability and Accountability Act. Pub. L. No. 104–191, § 264, 110 Stat.1936
10. Health Level 7 Standards. Health Level 7 International (2022). http://www.hl7.org/
11. Jacobs, M.L., Clawson, J., Mynatt, E.D.: My journey compass: a preliminary investigation of a mobile tool for cancer patients. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 663–672. Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2556288.2557194
12. Kessel, K.A., et al.: Mobile health in oncology: a patient survey about app-assisted cancer care. JMIR mHealth uHealth. **5**, e81 (2017). https://doi.org/10.2196/mhealth.7689
13. Low, C.A.: Harnessing consumer smartphone and wearable sensors for clinical cancer research. NPJ Digit. Med. **3**, 1–7 (2020). https://doi.org/10.1038/s41746-020-00351-x
14. Lung cancer statistics: How common is lung cancer? American Cancer Society (2022). https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html
15. Mauro, J., Mathews, K.B., Sredzinski, E.S.: Effect of a smart pill bottle and pharmacist intervention on medication adherence in patients with multiple myeloma new to Lenalidomide therapy. JMCP **25**, 1244–1254 (2019). https://doi.org/10.18553/jmcp.2019.25.11.1244
16. McGregor, B.A., Vidal, G.A., Shah, S.A., Mitchell, J.D., Hendifar, A.E.: Remote oncology care: review of current technology and future directions. Cureus. **12**, e10156 (2020). https://doi.org/10.7759/cureus.10156
17. Ni, X., et al.: Development of mobile health–based self-management support for patients with lung cancer: a stepwise approach. Nurs. Open **9**, 1612–1624 (2022). https://doi.org/10.1002/nop2.1185

18. Panayi, N.D., Mars, M.M., Burd, R.: The promise of digital (mobile) health in cancer prevention and treatment. Future Oncol. **9**, 613–617 (2013). https://doi.org/10.2217/fon.13.42
19. Pater, J., Owens, S., Farmer, S., Mynatt, E., Fain, B.: Addressing medication adherence technology needs in an aging population. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, pp. 58–67. ACM, Barcelona Spain (2017). https://doi.org/10.1145/3154862.3154872
20. Pereira-Salgado, A., et al.: Mobile health intervention to increase oral cancer therapy adherence in patients with chronic myeloid leukemia (the REMIND system): clinical feasibility and acceptability assessment. JMIR mMhealth uUhealth **5**, e184 (2017). https://doi.org/10.2196/mhealth.8349
21. Schmalz, O., et al.: Digital monitoring and management of patients with advanced or metastatic non-small cell lung cancer treated with cancer immunotherapy and its impact on quality of clinical care: interview and survey study among health care professionals and patients. J. Med. Internet Res. **22**, e18655 (2020). https://doi.org/10.2196/18655
22. Shiffman, S., Stone, A.A., Hufford, M.R.: Ecological momentary assessment. Annu. Rev. Clin. Psychol. **4**, 1–32 (2008). https://doi.org/10.1146/annurev.clinpsy.3.022806.091415
23. Skrabal Ross, X., Gunn, K.M., Patterson, P., Olver, I.: Development of a smartphone program to support adherence to oral chemotherapy in people with cancer. Patient Prefer. Adherence. **13**, 2207–2215 (2019). https://doi.org/10.2147/PPA.S225175
24. Theodos, K., Sittig, S.: Health information privacy laws in the digital age: HIPAA doesn't apply. Perspect. Health Inf. Manag. **18**, 1l (2021)
25. Toscos, T., et al.: Medication adherence for atrial fibrillation patients: triangulating measures from a smart pill bottle, e-prescribing software, and patient communication through the electronic health record. JAMIA Open **3**, 233–242 (2020). https://doi.org/10.1093/jamiaopen/ooaa007
26. Wood, L.: A review on adherence management in patients on oral cancer therapies. Eur. J. Oncol. Nurs. **16**, 432–438 (2012). https://doi.org/10.1016/j.ejon.2011.10.002
27. Xiong, H., Huang, Y., Barnes, L.E., Gerber, M.S.: Sensus: a cross-platform, general-purpose system for mobile crowdsensing in human-subject studies. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp 2016, pp. 415–426. ACM Press, Heidelberg, Germany (2016).https://doi.org/10.1145/2971648.2971711

Exploring the Design Space of Technological Interventions for Menopause: A Systematic Review

Kamala Payyapilly Thiruvenkatanathan(✉), Shaowen Bardzell, and Jeffrey Bardzell

Pennsylvania State University, University Park, PA 16801, USA
kamala.pt@psu.edu

Abstract. Menopause is a phase in a woman's lifecycle, which is considered to have occurred when a woman does not have a menstrual period for 12 consecutive months. Despite being perceived just as a biological phenomenon, a woman experiences several psychosocial symptoms along the way as she transitions into menopause, which begin well before the cessation of menstruation and sometimes continue to exist even after the onset of menopause, making it a 2-to-7-year journey. Ongoing research on women's health in the Human Computer Interaction (HCI) research community has led to an increasing number of works focusing on the intervention of technology in women's health, including a rising interest in designing technological interventions for menopause. With approximately 6000 women reaching menopause daily in the US, there is a need to understand the design space of technological interventions for menopause by surveying prior studies, to eventually contribute towards designing technologies for menopausal women. This paper presents a systematic review of prior studies on technological interventions for menopause. The aim of the review is to (1) Understand how prior studies have approached the design of technological interventions for menopause (2) Identify the technological features and goals of the interventions proposed by prior studies (3) Identify the symptoms (physiological/psychosocial) being addressed by the proposed interventions. A systematic review of 12 papers collected from the ACM Digital Library highlights the characteristics of prior studies on technological interventions for menopause, such as type of study, study design and interventions that are discussed in the study. Based on the findings, we discuss aspects that were comprehensively studied, potential design implications for interventions for menopause along with limitations of the current study and opportunities for future research on technological interventions for menopause.

Keywords: Menopause · Women's health · Human Computer Interaction

1 Introduction

Menopause is a phase in a woman's lifecycle, which is considered to have occurred when a woman does not have a menstrual period for 12 consecutive months [15]. Research has shown that about 6000 people reach menopause daily in the US [2], usually between the

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 584–599, 2023.
https://doi.org/10.1007/978-3-031-34586-9_39

ages of 45 to 55. Common symptoms associated with menopause include hot flushes, night sweats, reduced sex drive and headaches [7] with early research addressing the most common symptom of hot flush using hormone replacement therapy [15]. Despite being perceived just as a biological phenomenon, a woman experiences several psychosocial symptoms along the way as she transitions into menopause. Menopausal symptoms begin well before the cessation of menstruation and sometimes continue to exist even after the onset of menopause, making it a 2-to-7-year journey before reaching menopause [2]. This has led to the division of menopause into stages, with most of the symptoms occurring during the menopausal transition stage which is the phase between the onset of menstrual irregularities and menopause [15]. Research aiming to understand technology's support for menopausal women have claimed the need for technologies to facilitate social support [12] and to enable communication among women and Health Care Practitioners (HCPs), along with sharing information regarding menopause and self-care strategies [2]. As a result, menopause has been a growing area of interest for technological interventions [11].

Technological interventions have been influencing women's experiences with their bodies and health. With women's health research gaining traction in the HCI research community, an increasing number of works focus on designing and evaluating technological interventions for several women's health issues. Much of the works focus on menstrual health [8, 9] and maternal health [3, 16, 20, 25] followed by menopause [14, 19] and vaginal health [1]. Menstrual trackers are the most widely designed and evaluated menstrual health intervention, with works examining menstrual tracking applications to understand women's menstrual cycle tracking practices [8] and designing better menstrual tracking systems [9]. Interventions for maternal health include designing for breastfeeding [3], designing mobile health applications to support women during pregnancy [16, 25] and designing for emotional well-being of pregnant women [20]. Despite being fairly recent, works on technological interventions for menopause include proposing mobile health (mHealth) applications for menopause such as a menopausal period tracking system [14] and a persuasive coaching application for self-care during menopause [19]. Other forms of technological interventions proposed for menopause include wearables such as smart cooldown bra [23] and smart spaces using ubiquitous computing [4] among others.

With a growing interest in designing technological interventions for menopause, there exists a need to systematically review related literature in the HCI research community, aimed towards understanding the landscape of technological interventions designed or proposed for menopause and identifying opportunities for designing better technological interventions for menopausal women. While prior works within the HCI research community include systematic literature reviews of technological interventions for children with special needs [5] and health technologies for families [17], there is a dearth of works that systematically review literature on technological interventions for menopause. Informed by the body of literature that showcases the value of a systematic review, we propose a review of prior studies that combine menopause and technological interventions, in order to explore the design space of technological interventions for menopause. In this paper, we present findings from the review of 12 papers (the process of identifying this collection of 12 papers is detailed below) focused on technological interventions for menopause, systematically collected from the ACM Digital Library. The contribution of this systematic review is to characterize prior studies on technological interventions for menopause by identifying the technologies used, their features, goals of interventions (particularly the type of menopausal symptom being addressed) as well as methodologies adopted during the design process. The broader aim is to showcase how menopause is framed by prior studies proposing or designing technological interventions, contributing towards the mitigation of medicalization of menopause [4, 6].

2 Methods

The systematic review of technological interventions for menopause was a four-step process, beginning with a comprehensive search for related literature in the ACM Digital Library database followed by screening of the collected data to finalize the corpus for analysis. The following criteria were key to choosing a paper during the screening process (1) The paper must target menopausal women (2) The paper must include a design or proposal of technological intervention for menopause (3) If no technological intervention was designed or proposed, the paper must include implications or guidelines for designing technological interventions for menopause. The data collection process involved reviewing the finalized papers and collecting information particularly related to the study methodology, context and technological intervention being designed or proposed. The final step involved analyzing the data aimed towards answering the research questions (RQ). The following research questions guided this systematic review, during the collection and assessment of characteristics of prior studies on technological interventions for menopause:

- RQ1: What are the types of technological interventions designed or proposed for menopausal women?
- RQ2: What symptoms (physiological/psychosocial) have been explored in the design of technological interventions for menopausal women?
- RQ3: How have prior studies approached the design of technological interventions for menopause?

2.1 Database Search

As technological interventions for menopause designed or proposed within the HCI research community was the focus of this review, ACM Digital Library was chosen as the relevant database. With interventions for menopause falling within the broader area of interventions for women's health, a quick search was performed to assess the amount of data available within the database. The search targeting women's health and technological interventions within the ACM Digital Library - Full Text Collection resulted in a smaller corpus (78 publications falling between the years 2011 and 2022). For a comprehensive review, the search was expanded to ACM Guide to Computing Literature which resulted in a slightly larger corpus (91 publications falling between the years 2007 and 2022). As a result, the larger ACM Guide to Computing Literature database was used to search for literature related to technological interventions for menopause.

The primary search consisted of an iteratively developed boolean search string, containing a total of 18 terms combined with AND/OR operators, distributed across the key categories of menopause and technological interventions. For instance, terms related to menopause included “menopausal women”, “midlife women” and “perimenopause” to name a few. Whereas terms related to technological interventions included “digital health”, “health technologies” and “FemTech”. Since the corpus related to the broader category of women’s health was small to begin with, no constraints were imposed related to the publication years or publication venue. The resulting corpus consisted of 65 publications, falling between the years 2005 and 2022. Since prior review of literature related to women’s health interventions revealed the frequent use of mobile health (mHealth) as an intervention technology, a secondary search was done, specifically looking for mHealth technological interventions for menopause. The search string included the previous terms related to menopause and instead of terms related to technological interventions, specific terms related to mHealth such as “mobile health”, “mobile app” and “mobile health application” were used in the boolean search string. To avoid repetition, the search excluded publications related to menopause and technological interventions. The resulting corpus from the secondary search included 34 publications (after removing duplicates). At the end of the database search, a total of 99 publications were moved to the next step of title, abstract and full text screening.

2.2 Screening Process

The first step in the screening process was to review the total set of 99 publications for availability of full text. This was conducted since a comprehensive search within the ACM Guide to Computing Literature database resulted in extended abstracts being included. Having excluded 6 publications, owing to the unavailability of full text, 93 publications were moved to the title screening step. Publication titles were reviewed for relevance to menopause and technological interventions. Since the corpus was comparatively small, only the completely irrelevant titles, that were neither related to technological interventions nor narrowly related to menopause, were excluded during the screening process. This step led to 82 publications screened for their abstracts which involved reading abstracts of the chosen publications. The publications whose abstracts were directly related to menopause or designing technological interventions for menopause were moved to the full text screening step. While abstracts completely unrelated to menopause resulted in the publications being excluded, several abstracts addressed women’s health issues more broadly, thereby showing potential to be related to menopause. This resulted in publications related to broader women’s health issues also being moved to the next step, resulting in a total of 29 publications being included in the full text screening process. A comprehensive review of the full text of the chosen publications resulted in the final corpus containing 12 publications, ranging between the years 2015 and 2021. Figure 1 shows the process followed during the systematic review that led to the final corpus.

2.3 Data Collection

To curate the corpus for data analysis, each publication was reviewed, and key pieces of information related to the study were extracted. This included details regarding the type of study, study design, empirical data being used by the study (along with participant

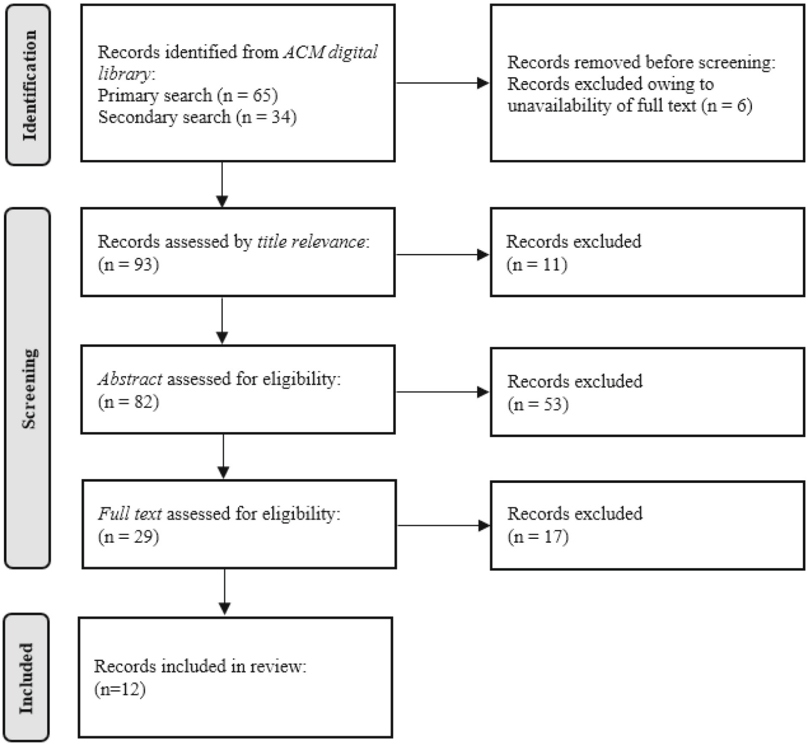

Identification
Records identified from ACM digital library:
Primary search (n = 65)
Secondary search (n = 34)
Records removed before screening:
Records excluded owing to unavailability of full text (n = 6)
Screening
Records assessed by title relevance:
(n = 93)
Records excluded
(n = 11)
Abstract assessed for eligibility:
(n = 82)
Records excluded
(n = 53)
Full text assessed for eligibility:
(n = 29)
Records excluded
(n = 17)
Included
Records included in review:
(n=12)

Fig. 1. PRISMA flow diagram showing the number of publications at each stage of the systematic review

demographics) and details regarding the technological interventions being designed or proposed. Publications that did not include design of technological interventions, empirical data or those that included just implications for designing technological interventions for menopause, were flagged for exception despite their key information being recorded. Table 1 shows the corpus of papers on technological interventions for menopause, sorted in ascending order by year, that this literature review is based on.

2.4 Data Analysis

Data analysis involved analyzing the key pieces of information curated during the data collection process and associating them with the research questions, in search of responses. Since the corpus was relatively small, the analysis focused on the characterization of studies discussed in each paper rather than trying to analyze the corpus quantitatively or qualitatively. The purpose of analysis was to understand the landscape of ongoing research on technological interventions for menopause and generate preliminary implications based on the results and guidelines disclosed by the studies being reviewed. With the broader goal aimed towards understanding how the prior studies and

Table 1. The corpus of papers on technological interventions for menopause.

Paper	Author(s)	Year	Technological Intervention
Designing MHealth Intervention for Women in Menopausal Period	Lee et. al [14]	2015	Mobile health (mHealth) for menopausal women's wellness
Understanding Women's Needs in Menopause for Development of MHealth	Lee et. al [13]	2015	Mobile health (mHealth) for menopausal women's wellness
Participatory User Requirements Elicitation for Personal Menopause App	Trujillo & Buzzi [21]	2016	Mobile application for personalized coaching system for menopausal women
Persuasive Design of a Mobile Coaching App to Encourage a Healthy Lifestyle during Menopause	Senette et al. [19]	2018	Mobile application for personalized coaching system for menopausal women
Towards a Fuzzy Rule-Based Systems Approach for Adaptive Interventions in Menopause Self-Care	Trujillo & Buzzi [22]	2018	Mobile application for personalized self-care
Parting the Red Sea: Sociotechnical Systems and Lived Experiences of Menopause	Lazar et al. [12]	2019	No technological intervention design is being proposed
Inaction as a Design Decision: Reflections on Not Designing Self-Tracking Tools for Menopause	Homewood [10]	2019	No technological intervention design is being proposed
(Re-)Framing Menopause Experiences for HCI and Design	Bardzell et al., [4]	2019	Tracking and relieving stress balls, Smart vibrator, Smart mirror, Smart spaces using ubiquitous computing
HCI and Menopause: Designing With and Around the Aging Body	Tutia et al. [23]	2019	Wearable smart cool down bra, Mobile application for [23] self-tracking, Interactive story map, Adaptive gathering tool
Resisting the Medicalisation of Menopause: Reclaiming the Body through Design	Ciolfi Felice et al. [6]	2021	Smart textile wearable
Utopian Futures for Sexuality, Aging, and Design	Schulte et al. [18]	2021	Wearable smart lingerie

(*continued*)

Table 1. (*continued*)

Paper	Author(s)	Year	Technological Intervention
Designing Interactive Technological Interventions for Menopausal Women	Warke [24]	2021	Web based menopausal digital diary, Wearable interactive jewelry

the designed or proposed technological interventions have framed menopause. Following section presents the findings from the analysis of the collected corpus, sectioned based on key information used for assessing each study.

3 Findings

This section reports findings from the systematic review of 12 papers focused on technological interventions for menopause. The sections are categorized based on the pieces of information that played a key role in characterizing each study against the research questions.

3.1 Type of Study

Following others [17], type of study in each paper was classified as formative, evaluative or design proposal. Studies were considered formative if they discussed research conducted to elicit design implications or guidelines for designing technological interventions for menopause. Evaluative studies were those that evaluate and report the impact of new or existing technological interventions for menopause. Papers were considered to contain a design proposal study if they propose a novel design of a technological intervention for menopause. Most of the papers reported a combination of studies, such as formative and design proposals or design proposals and evaluative.

Among the 12 papers, only 4 papers discussed an exclusive type of study, with 3 of them being formative suggesting guidelines for designing interventions for menopause based on understanding menopausal women's experiences [12, 14, 18] while 1 of them containing design proposals [24] for technological interventions for menopause. A total of 4 papers discussed studies that were both formative and design proposal, among which 1 paper contained formative guidelines followed by designs proposed by the researchers [13], 1 paper contained formative guidelines but design proposed by participants of research [6], 1 paper discussed formative guidelines followed by design provocations rather than proposals [23] and 1 paper contained speculative design proposals followed by guidelines for designing interventions to enhance menopausal women's experiences [4]. Among the 12 papers, 2 of them included a design proposal and an evaluative study containing a straightforward design proposal of an mHealth intervention followed by evaluation of the same [19, 22]. There were 2 papers that were within an exception category of formative and evaluative studies, with 1 containing formative guidelines and use case for interventions for menopause which were evaluated by participants [21] and the other taking a reflective approach in evaluating existing interventions for menopause and proposing inaction as a guideline for designing menopausal interventions [10]. Note that some of the design proposals looked beyond using computational technology, as they included design of textiles, services, and speculative scenarios among others.

3.2 Study Design

This section reports details regarding the study discussed in the 12 papers being reviewed. Participant size and demographic characteristics are reported wherever available, with participant involvement denoting the contribution that participants have towards the design of technological interventions for menopause. Context of study and research methodologies are included to curate the approaches that prior works have taken in designing technological interventions for menopause.

Participant Size and Demographic Characteristics
Among the 12 papers reviewed, only 8 of them reported some form of empirical data that involved participants. Within the 8 studies containing data related to participants, 2 pairs of papers and a group of 3 papers, were related. To elaborate, papers [13, 14] were both related and works of the same authors, with [14] involving a total of 13 participants including middle aged women between 45 and 60 and their family members, located at South Korea, to curate guidelines for designing technological interventions for menopause based on participant requirements and [13] proposing prototype of an mHealth intervention for menopausal women, designed based on guidelines gathered during the study discussed in [14]. Similarly, papers [4, 12] were related, with [12] reporting findings from analysis of empirical data scraped from an online forum, while [4] used the same to propose speculative interventions for menopause. The papers [4, 12], despite not having direct participant involvement, contained analysis of 300 discussion threads and 2065 corresponding comments, from covering approximately 72% of a Subreddit menopause forum corpus. The papers [19, 21, 22] were all related to the same study (divided into four phases) based off of Tuscan region in Italy, with [21] eliciting initial requirements for a mobile application for menopause gathered from 26 women experiencing pre-menopause or menopause (the first two phases), [19] designing a prototype based on the requirements gathered and evaluating the same with 14 women (the third phase) and [22] developing the technical model for the mobile application for menopause, involving a total of 34 participants (missing demographic information) during the evaluation of model variables (fourth phase). Among the other 5 papers, only 3 papers reported on some form of participant data. The paper [23] reported on a study involving 17 cisgender women, recruited from social media groups related to menopause, towards understanding participant requirements in designing technological interventions for menopause. The paper [6] insisted on the value of participants being active actors during the design process, by involving a total of 12 women between the ages 44 and 58 from one or more of the countries including Argentina, Sweden, Finland, and France, across two phases, eventually leading to participants proposing designs of interventions for menopause. The opinion paper [18] summarized a workshop exploring the role of technology to support aging women with their intimate life, that involved active conversations among participants and researchers, coming from a Western European perspective, but the paper did not disclose the exact participant size. Further, papers

[10, 24] had no form of participant data with [10] taking an autoethnographic, reflective approach towards designing for menopause whereas [24] contained design proposals without concrete empirical data.

Research Methodology

Research methodology adopted by each study was included in the review, to understand how prior studies have approached the design of technological interventions for menopause. From the review, it was evident that most of the studies took the efforts to actively engage with participants' experiences including eliciting their requirements and ideas during the design and evaluation of interventions. For instance, papers [18, 23] discussed studies that adopted a participatory design approach to encourage participants to actively communicate their experiences and ideas using a speculative narrative. Another form of participatory design method discussed, related to participants designing interventions with materials provided during the study [6]. Some other common methods adopted by several studies were focus group interviews [14, 21, 23] and semi-structured interviews [6, 14, 23], both aimed towards gathering firsthand qualitative data from participants' experiences. The papers [4, 12] adopted a theoretical approach towards analyzing secondary data collected from scraping, along with [4] using a speculative design methodology to propose design interventions for menopause. Papers that were exceptions involved unique methodologies including [19] that described an evaluative study using thinking aloud protocol and [10] where the researcher reflected based on the evaluation of existing interventions for menopause, to eventually decide not to design technological interventions for menopause.

Participant Involvement and Context of Study

All studies that involved participatory design as a research methodology [6, 18] showcased ample involvement, with participants actively contributing towards the formulation of design implications or towards the design of technological interventions for menopause. Participant involvement was also observed in studies involving focus group interviews [14, 21, 23] as well as semi-structured interviews [6, 14, 23], as firsthand experiential information from participants contributed towards the design of interventions for menopause. Evaluative studies [19, 21, 22] also showcased participant involvement with participants actively involved in evaluating use cases [21], evaluating prototypes [19] where participants expressed their concerns related to data privacy and evaluating model variables [22]. Exception papers related to participant involvement included paper [12] that had secondary data collected from an online forum and papers [10, 24] that had no form of participant data.

With respect to the context, studies involving participatory design, focus group interviews as well as evaluations were conducted in a design lab setting or researchers' institution with only one study [6] reporting that the semi-structured interviews were conducted including at the participants' homes or workplaces. The context of the study was included in the review to assess how much of menopausal women's contextual information was included during the design of technological interventions for menopause.

An overall observation from the review of participant data and study design showed that efforts were being made to understand menopausal women's experiences, by deploying participatory design research methods, either through firsthand participation or through secondary data, during the design of technological interventions for menopause. The studies reviewed so far, however, suffered from the fact that interventions for menopause were predominantly west centered, revealing opportunities to design technological interventions for menopausal women in the global south.

3.3 Technological Intervention

This section reports details regarding the technological interventions designed or proposed by the 12 papers being reviewed. Details regarding the technological interventions including the type of technology and features are reported wherever available. The subsection on goal of intervention is aimed towards assessing the symptoms being addressed by the designer or proposed intervention, with a broader goal of understanding how the interventions frame menopause.

Type of Technology and Features

Among the 12 papers reviewed, only 8 of them involved some form of technological intervention being designed or proposed. With papers [13, 14] being related, the intervention being proposed was an mHealth intervention focused on menopausal women's wellness, that allows menopausal women to record their menstrual cycle, provide personalized information from health professionals and social support. The papers [19, 21, 22] were also related to the same study, with the proposed intervention being a persuasive mobile application that acts as a mobile coaching system for menopausal women, automatically adapting to their personalized health needs. The paper [23] proposed design provocations including wearable technologies such as smart cool down bra and mobile application for self-tracking menopausal symptoms and sharing educational information. Despite being a research proposal with no empirical data, the paper [24] proposed a web based technological intervention, the menopause digital diary, to record menopausal women's daily personal stories. The paper also included a design proposal for an interactive jewelry that takes the form of a wearable, for tracking quantifiable menopausal symptoms. The paper [4] included proposal of smart vibrator, smart mirror and smart spaces designed using ubiquitous computing, to cater to menopausal women's needs. Papers also included design proposals for interventions for menopause that are beyond technology, details of which are discussed in the following sections.

Goal of Intervention

The review included the goal of technological interventions being designed or proposed, in order to assess how prior studies have framed menopause, particularly based on the symptoms such as physiological or psychosocial. As stated earlier, menopause has been often perceived as a biological phenomenon with research proposing hormone replacement therapy to address the most common menopausal symptom of hot flush [15]. The assessment of the goals of technological interventions was approached with the presumption that works continue to reduce menopause to a set of biological symptoms. However, our presumptions were proven wrong when most of the studies showcased the efforts taken in addressing menopausal women's overall well-being. This was particularly the case with the papers [19, 21, 22] that focused on designing a mobile coaching application aimed towards self-care for menopausal women. The same was the case with the papers [13, 14] whose aim was to design mHealth intervention for menopausal wellness, though there was a lack of clarity on the definition of menopausal wellness. In the paper [4] the authors clearly stated their belief that menopause is beyond physiological symptoms and the same was evident in their design proposals that were aimed towards addressing menopausal women's overall well-being. Despite proposing a wearable for the specific symptom of hot flush, in the paper [23] the authors stated that the focus of their design interventions is to design with and around menopausal women's overall well-being. The same was the case in the papers [6, 18] where the proposed designs seemed to be related to overall well-being of aging women and enhancing their menopausal experiences. In the case of [24], the proposed interventions were aimed towards collecting menopausal women's experiential and quantifiable data, to serve as a repository for women to understand their own menopausal experiences and eventually enhance their quality of life.

An overall observation was that mHealth technology was the most opted form of technological intervention with proposed designs taking the form of mobile applications. Understanding the goals of interventions revealed that efforts are being taken to mitigate the medicalization of menopause, by designing for menopausal women's overall well-being rather than a specific physiological or psychosocial symptom. The challenge, however, was the lack of evaluation of the proposed designs to understand their effectiveness in providing delightful experiences for menopausal women. Further, most of the studies seemed to lack consideration of context and did not explicitly discuss how the proposed interventions can measure and enhance menopausal women's contextual experiences.

3.4 Beyond Technology

This section was a result of the systematic review as some of the papers included design proposals for interventions for menopause, that are beyond computational technology. To elaborate, the current review approached technological interventions as those that involve some form of computational technology including mHealth, smart wearables and telehealth, among others. However, the collected corpus not only contained technological interventions for menopause but also included other forms of design interventions such as design of textiles, services, and scenarios, all contributing towards a better menopausal experience. For instance, [4] discussed speculative interventions such as a menopause lifestyle brand that provides products and services aimed at the new freedom available menopausal women. In [6], the participatory design study resulted in the design of cocoon and spike mat, both being made from textile materials inspired by Soma design. The resulting narratives from a participatory workshop summarized in [18] included a sexual care package subscription and lingerie suitable for women using urinary incontinence pads, both aimed towards enhancing aging women's bodily experiences. Presence of other forms of design interventions for menopause in the corpus opened opportunities for design implications that look beyond technological interventions for menopause. Further, it also exposed a gap in the current review, highlighting opportunities for future work.

4 Discussion

The synthesis of findings from the review of 12 papers on technological interventions for menopause serves as a starting point for initiating conversations on the role of HCI in designing interventions for menopause and eventually contributing towards mitigating medicalization of menopause. In the following section, we reflect on how the assessment of the corpus addresses the proposed research questions, by characterizing the technological interventions being discussed. This section also sheds light on how such interventions construct menopause, exposing the existing limitations and discussing implications and opportunities for future research that contributes towards de-medicalization of menopause. The proposed implications are nascent owing to the size of the corpus, which in itself reveal an opportunity - of the need for HCI to intervene more actively in designing interventions for menopause.

4.1 Characterizing Technological Interventions

The responses to the research questions guided the process of characterizing technological interventions for menopause focusing particularly on the aspect of menopause that is being addressed by the intervention (such as a symptom) leading to understanding how the intervention operationalizes menopause. We approached the review with an a priori hypothesis that studies continue to reduce menopause to a set of biological symptoms. The hypothesis was falsified, when the assessment of studies showcased the efforts taken by interventions to address menopausal women's overall well-being rather than focusing on a specific physiological or psychosocial symptom. However, interrogating each of the technological interventions being designed revealed menopausal women's health data that is being captured and processed, showcasing their limitations. For instance, [13, 14] propose the design of an mHealth intervention that support menopausal women's overall wellness, by capturing data related to menopausal women's period cycle alongside personal demographic information. The proposed intervention is claimed to share personalized menopause related information by gauging the woman's menopausal phase. Additionally, the intervention is claimed to push messages with exercise suggestions, thereby constructing menopause as a biological condition. A similar approach is taken in [19, 21, 22] where the authors propose the design of a persuasive menopause mobile application that supports self-care. Analyzing the prototype, however, revealed that the application captures menopausal women's physiological symptoms such as hot flush, osteoporosis etc. alongside diet and steps walked, with the goal to persuade users into making behavioral changes, leading to the reduction of cardiovascular risks caused by menopause. By focusing on an effect of menopause that is clearly physiological, the aforementioned study continues to construct menopause as a biological condition. Several other studies, while speculatively proposing technological interventions, tended to focus on either a specific physiological symptom such as hot flush [23] or quantifying symptom related data using sensors and bio signals [24], thereby reducing menopause to a biological construct. Very few studies that proposed the design of services and non-technological interventions for menopause [4, 18] looked beyond the physiological symptoms of menopause.

In summary, a closer look at the interventions revealed that, while efforts are being taken to support menopausal women's holistic wellbeing, there remains a gap in translating those efforts into design of tangible technological interventions. Additionally, it was evident that non-technological interventions focused more on designing interventions for menopausal women's overall wellness.

4.2 Construction of Menopause

We approached the analysis with the presumption that menopause has been overmedicalized. However, several studies proved otherwise, with their discourses centered around designing technological interventions for menopausal women's holistic wellbeing. Comparing the discourse of the studies being reviewed alongside the technological interventions being proposed, including the data being collected and its functionalities, however, revealed a gap in the discourse being translated into the technological intervention being proposed. We observed that menopause continued to be operationalized as a biological construct, with interventions capturing quantifiable data and taking an information processing approach. While the overall analysis showed that studies are moving away from medicalization of menopause, towards holistic wellbeing, it was evident that the claims were nascent with the technological interventions continuing to quantify menopausal experiences. The studies that proposed non-technological interventions were more aligned towards supporting menopausal women's holistic wellbeing but fell short of being implemented and evaluated and remained speculative. Additionally, the review showed that studies on interventions for menopause were predominantly west centered, revealing opportunities to design technological interventions for menopausal women in the global south.

In summary, the assessment of the interventions discussed within the corpus showed that efforts are being taken to mitigate the medicalization of menopause, but the technological interventions being proposed lacked clarity on how the efforts were translated to support menopausal women's holistic wellbeing.

4.3 Towards De-medicalization of Menopause

The assessment of studies proposing technological interventions for menopause, revealed opportunities to look beyond physiological symptoms and quantifiable menopausal experiences. Further, the synthesis revealed design interventions for menopause beyond computational technology, through design proposals of textiles, services, and scenarios, all of which were aimed towards addressing menopausal women's holistic well-being. As an attempt to contribute towards de-medicalization of menopause, a potential design implication would be to propose system designs that combine emergent computational technologies along with non-technological interventions such as services [4], soma design inspired products [18], care packages [18] etc. aimed towards not only capturing quantifiable data but also providing enhanced experiences for menopausal women's overall well-being. This implication was inspired by the design proposal discussed in [4] where the authors propose the design of Menobuddy, a traditional doll combining computing capabilities to record and playback menopausal experiences in the form of stories.

In order to contribute towards the de-medicalization of menopause, we suggest that technological interventions be designed acknowledging the "entanglements of the physical and psychosocial" [6] experiences during menopause.

5 Limitations and Future Work

The aim of this paper was to characterize prior studies on technological interventions for menopause by identifying the symptoms (physiological/psychosocial) being addressed and showcasing how menopause is framed, eventually contributing towards the mitigation of medicalization of menopause. A systematic literature review was adopted as the methodology, to survey the landscape of studies on technological interventions for menopause within the HCI research community.

The database searched to create the literature corpus was the ACM Digital Library. Despite being recent, designing interventions for menopause have been an emerging area of interest in various fields. However, this review was focused only on interventions proposed by the HCI research community, with an underlying assumption that the ACM Digital Library is the optimal database for a comprehensive review. The resulting size of the corpus, however, revealed the need for looking beyond the ACM Digital Library, to include other databases such as Google Scholar, to expand the size of the corpus. The primary database search focused on technological interventions for menopause with the secondary search narrowing down to mHealth interventions. However, there could have been other forms of technological interventions for menopause, which were not captured by the current review. Furthermore, the review was intended to survey the landscape of technological interventions for menopause by characterizing studies related to the same, rather than assessing the quality of each study. As a follow on, review of the corpus also revealed other forms of design interventions for menopause, that are not necessarily technological, which are also not captured in the current review.

Menopausal conditions are often experienced alongside other related health conditions, owing to the woman's age. This means that there could be technological interventions not necessarily designed for menopause but cater to menopausal women's health needs. Since the current review focused exclusively technological interventions for menopause, a potential future work can be aimed towards expanding the corpus to include interventions that directly or indirectly cater to menopausal women's health needs, to eventually propose technological interventions that cater not only to menopause but potentially to other health conditions that women experience alongside menopause, designing for those experiences as a whole.

6 Conclusion

This review reports a systematic synthesis of 12 papers in HCI literature focused on technological interventions for menopause, aimed towards exploring the design space for creating interventions for menopause. The findings revealed that designing technological interventions for menopause has been a rising area of interest, with studies being fairly recent (2015 to 2021) but most of them aiming to address menopausal women's overall well-being rather than reducing menopause to a set of biological symptoms, by involving women during the research and design process. The findings also revealed that interventions for menopause can expand beyond computational technology, to cater to menopausal women's health needs. Based on the findings, we unpack design implications for menopause and recommend areas where HCI might be able to intervene.

References

1. Almeida, T., Comber, R., Wood, G., Saraf, D., Balaam, M.: On looking at the vagina through labella. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 1810–1821 (2016). https://doi.org/10.1145/2858036.2858119
2. Backonja, U., Taylor-Swanson, L., Miller, A.D., Jung, S.-H., Haldar, S., Woods, N.F.: There's a problem, now what's the solution?: Suggestions for technologies to support the menopausal transition from individuals experiencing menopause and healthcare practitioners. J. Am. Med. Inform. Assoc. **28**(2), 209–221 (2021). https://doi.org/10.1093/jamia/ocaa178
3. Balaam, M., Comber, R., Jenkins, E., Sutton, S., Garbett, A.: FeedFinder: a location-mapping mobile application for breastfeeding women. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems – CHI'15, pp. 1709–1718 (2015). https://doi.org/10.1145/2702123.2702328
4. Bardzell, J., Bardzell, S., Lazar, A., Su, N.M.: (Re-)Framing menopause experiences for HCI and design. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2019). https://doi.org/10.1145/3290605.3300345
5. Baykal, G.E., Van Mechelen, M., Eriksson, E.: Collaborative technologies for children with special needs: a systematic literature review. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2020). https://doi.org/10.1145/3313831.3376291
6. Ciolfi Felice, M., Søndergaard, M.L.J., Balaam, M.: Resisting the medicalisation of menopause: reclaiming the body through design. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (2021). https://doi.org/10.1145/3411764.3445153
7. Cronin, C., Hungerford, C., Wilson, R.L.: Using digital health technologies to manage the psychosocial symptoms of menopause in the workplace: a narrative literature review. Issues Ment. Health Nurs. **42**(6), 541–548 (2021). https://doi.org/10.1080/01612840.2020.1827101
8. Epstein, D.A., et al.: Examining menstrual tracking to inform the design of personal informatics tools. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 6876–6888 (2017). https://doi.org/10.1145/3025453.3025635
9. Fox, S., Howell, N., Wong, R., Spektor, F.: Vivewell: speculating near-future menstrual tracking through current data practices. In: Proceedings of the 2019 on Designing Interactive Systems Conference, pp. 541–552 (2019). https://doi.org/10.1145/3322276.3323695
10. Homewood, S.: Inaction as a design decision: reflections on not designing self-tracking tools for menopause. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2019). https://doi.org/10.1145/3290607.3310430
11. Kemble, E., Perez, L., Sartori, V., Tolub, G., Zheng, A.: Unlocking opportunities in women's healthcare. Healthcare Systems and Services: McKinsey & Company, 14 February 2022. https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/unlocking-opportunities-in-womens-healthcare
12. Lazar, A., Su, N.M., Bardzell, J., Bardzell, S.: Parting the red sea: sociotechnical systems and lived experiences of menopause. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–16 (2019). https://doi.org/10.1145/3290605.3300710

13. Lee, M., Koo, B., Jeong, H., Park, J., Cho, J., Cho, J.: Designing MHealth intervention for women in menopausal period. In: Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare, pp. 257–260 (2015a)
14. Lee, M., Koo, B., Jeong, H., Park, J., Cho, J., Cho, J.: Understanding women's needs in menopause for development of MHealth. In: Proceedings of the 2015 Workshop on Pervasive Wireless Healthcare, pp. 51–56 (2015b). https://doi.org/10.1145/2757290.2757295
15. Lund, K.J.: Menopause and the menopausal transition. Med. Clin. North Am. **92**(5), 1253–1271 (2008). https://doi.org/10.1016/j.mcna.2008.04.009
16. Sajjad, U.U., Shahid, S.: Baby+: a mobile application to support pregnant women in Pakistan. In: Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct - MobileHCI'16, pp. 667–674 (2016). https://doi.org/10.1145/2957265.2961856
17. Sandbulte, J., Byrd, K., Owens, R., Carroll, J.M.: Design space analysis of health technologies for families: a systematic review. In: Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare, pp. 21–37 (2020). https://doi.org/10.1145/3421937.3421988
18. Schulte, B., Søndergaard, M.L.J., Brankaert, R., Morrissey, K.: Utopian futures for sexuality, aging, and design. Interactions **28**(3), 6–8 (2021). https://doi.org/10.1145/3460204
19. Senette, C., Buzzi, M.C., Paratore, M.T., Trujillo, A.: Persuasive design of a mobile coaching app to encourage a healthy lifestyle during menopause. In: Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia, pp. 47–58 (2018). https://doi.org/10.1145/3282894.3282899
20. Svenningsen, I.K., Almeida, T.: Designing for the emotional pregnancy. Companion Publication of the 2020 ACM Designing Interactive Systems Conference, pp. 145–150 (2020). https://doi.org/10.1145/3393914.3395897
21. Trujillo, A., Buzzi, M.C.: Participatory user requirements elicitation for personal menopause app. In: Proceedings of the 9th Nordic Conference on Human-Computer Interaction (2016). https://doi.org/10.1145/2971485.2996737
22. Trujillo, A., Buzzi, M.C.: Towards a fuzzy rule-based systems approach for adaptive interventions in menopause self-care. Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, pp. 53–56 (2018). https://doi.org/10.1145/3213586.3226193
23. Tutia, A., Baljon, K., Vu, L., Rosner, D.K.: HCI and menopause: designing with and around the aging body. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–8 (2019). https://doi.org/10.1145/3290607.3299066
24. Warke, B.: Designing interactive technological interventions for menopausal women: designing and developing interactive technology tools to help aging women navigate information about stages of menopause to increase self-awareness of biopsychosocial changes and manage lifestyle for an improved quality of life. In: Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction (2021). https://doi.org/10.1145/3430524.3443692
25. Wierckx, A., Shahid, S., Al Mahmud, A.: Babywijzer: an application to support women during their pregnancy. In: Proceedings of the Extended Abstracts of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI EA '14, pp. 1333–1338 (2014). https://doi.org/10.1145/2559206.2581179

Personalized Healthcare

Understanding Barriers of Missing Data in Personal Informatics Systems

Nannan Wen[1](✉), Aditi Mallavarapu[2], Jacob Biehl[1], Erin Walker[3], and Dmitriy Babichenko[4]

[1] Department of Computer Science, School of Computing and Information, University of Pittsburgh, Pittsburgh, USA
{naw66,biehl}@pitt.edu
[2] Digital Promise Global and University of Pittsburgh, Pittsburgh, USA
amallavarapu@digitalpromise.org
[3] School of Computing and Information and Learning Research and Development Center, University of Pittsburgh, Pittsburgh, USA
eawalker@pitt.edu
[4] Department of Informatics and Networked Systems, School of Computing and Information, University of Pittsburgh, Pittsburgh, USA
dmb72@pitt.edu

Abstract. Advanced personal informatics (PI) tools enable users to collect and reflect on a wide range of personal data. Researchers consider missing data, or discontinuous (sparse) data caused by device malfunctions or human errors, an important barrier for adopting PI tools in their daily routines. While a lot is known about why missing data occurs, less is known about its impact on user reflection or how tools can be designed to mediate/reduce its negative effects in PI systems. In this work, we focused on exploring the importance and impact missing data has on user reflection and extracting insights to improve the design of PI reflection tools. We present a semi-structured interview to investigate the impact of missing data on users' daily usage on two user groups, trainees and maintainers. We then provide design implications for incorporating visualization of estimated data (synthetic data) in the reflection stage, as a potential solution to the missing data problem. In this work, we provided data-driven implications for the design of future PI tools to help users reflect upon and mitigate missing data in their tracking activities.

Keywords: Personal Informatics (PI) · missing data · visualization · synthetic data

1 Introduction

Personal Informatics (PI), also referred to as "living by numbers", "quantified self", "self-surveillance", "self-tracking", and "personal analytics", is defined as "those [tools] that help people collect personally relevant information for the

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 603–618, 2023.
https://doi.org/10.1007/978-3-031-34586-9_40

purpose of reflection and gaining self-knowledge" [20]. PI data have been used for tracking health through diet [1,35,36], exercise [23,24], changes in moods [2], sleep [25], menstrual cycles [14], and other physical and psychological activities. Extensive prior research has examined the use and utility of these technologies to improve behavior [6,21,27], understand differences in engagement and reflection on personal data [27], and build descriptive and predictive models of specific events and activities [13,20]. Research has also sought to expand what can be quantified, exploring new use domains [4,8,11,12,29]. In this work, we broadly refer to PI tools as the set of technologies and approaches that include wearable devices, systems, and applications that participants use to collect data on their personal behaviors and activities.

While there is great promise for PI-driven interventions, many practical obstacles impact users' ability to *collect* and *reflect* upon PI tools and associated data. For example, Li et al. [20] note three main barriers to data *collection*: 1) tools (devices) are not around when symptoms happen, 2) users forget to record activities, and 3) devices and applications lack necessary accuracy for helpful measurement. In *reflection*, the main obstacles include the sparsity of data and the ability to interpret data and resulting summaries, visualizations, and recommendations.

These obstacles have been shown to have a significant impact on adherence, interest, and trust in PI [10,13]. For example, Choe et.al. [6], who studied the practices of the Quantified Self movement, reported that participants valued tools (devices) that could capture comprehensive, granular information about their activities and expressed frustration when collected data was not accurate. Rapp et.al. [27] found that missing data might cause the users to mistake what activities were captured. Thomas et.al. [16], in their study of long-term use of wearable devices, describe that participants spent extra effort to ensure that they had their tools (devices) with them before their workout. They worried they would not get proper credit for their activities without the devices. Identifying incomplete data can also impact users' affect; Epstein et al. noted that participants felt guilty when a menstrual tracking application's interface noted missing data [12].

The literature, overall, notes that missing data plays a significant role in defining the value of PI [6,10,12,16,19,27], but few of them focused on resolving the effects of missing data. In this work, we conduct interviews in the hope of better understanding missing data's impact on the reflection stage and propose design implications to mediate its negative effect.

2 Related Work: Systems that Support Refection in PI

The most widely used definition of reflection in PI follows Schön's [31] reflection-in-action and reflection-on-action [3]. Systems that support reflection-in-action provide feedback during the activity. This concept is widely adopted in applications (e.g., [7,26,33]) that aim to promote physical health and encourage a healthy lifestyle using different methodologies. For example, commercial products like Fitbit and iWatch give real-time feedback when walking, running, or cycling. The Ally+ app [26] is an academic prototype that acts as a chat-based digital coach to deliver in-the-moment interventions to motivate participants to achieve their step goals. Systems supporting the concept of reflection-on-action usually provide feedback after the activity ends. Fish'n'Steps [22] is an example of such a system that promotes reflection-on-action; it is a game that links a player's daily step count to the growth of a virtual character, a fish, and a tank, to encourage physical activities. In the same line of work, the UbiFit [9] is another educational system designed to improve physical activity by using positive reinforcement based on past behaviors. In addition, Visualized Self [5] is a web-based system that supports deeper-level self-reflection through multiple data streams and visual data exploration using participants' historical data. Finally, Habito [17] is an android application that utilizes textual feedback on participants' activities to study how users engage with activity trackers.

While there are many different techniques to facilitate reflection, a specialized field of "personal visualizations" aims to present personally relevant information that promotes actionable insights and subsequent changes in behavior. Among those visualizations designed to increase awareness and encourage behavior change of self-trackers(e.g., [5,17,22]), many of them provided visualizations in the form of dashboards and supported simple interactions to explore the data, employing timeline metaphors to present events chronologically. For instance, Lifestreams [18] is an analytical tool to extract specific behavioral indicators and inferences from linear, interactive visualizations. Likewise, Moushhumi et al. [34] visualize time-series data to design just-in-time adaptive stress management interventions. These tools provided a valuable overview of trackers' activities.

Even though a considerable number of systems are available to aid data reflection, missing data was not considered a principle design element of their tools to the best of our knowledge. Nevertheless, it is well-recognized that missing data is an important barrier to PI. Two models have been proposed in the past two decades to study user behaviors through Personal Informatics systems. In the stage-based model, Li et.al. [20] divide the process of personal informatics into an iterative stage of preparation, collection, integration, reflection, and action. Epstein et.al. [13] grew their model by incorporating the perspective of lived informatics [28]. The interrelationship between stages in the lived informatics model is more complex than in the stage-based model due to the iterative nature between stages and substages. The lived-informatics model [28] includes the process of deciding, selecting, tracking & acting, and lapsing. Tracking & acting were further divided into an iteration of the collection, reflection, and integration process due to frequently switching and abandoning tools and periodically reviewing or reflecting on their data.

In Li's model, due to the cascaded dependency between the different stages, starting from the collection stage, missing and inaccurate data prevent people from going to the next stage or continuing the tracking activity. In the reflection stage, in both of this models [20,21], the discontinuous and missing data further prevent users from transitioning to the next stage, the action stage, through the lack of rich insight. In the lived-informatics model, Epstein et.al. [13] also identified that the reasons for stopping to track and starting again are the same barriers that Li et.al. [20] pinpointed. Even though researchers suggest tools should collect as much data as possible [20] to mediate the effect missing data has in the reflection stage, missing data are still present and causing problems.

3 Method

3.1 Research Design

Our study sought to understand the role of missing data and its subsequent impact on PI usage, utility, and related behaviors. Specifically, we sought to answer the following research questions: What are users' expectations of consistency and completeness in PI data? Does missing PI data conflict with these expectations? How can the design of PI tools be extended to help users reflect upon and mitigate missing data in their tracking activities?

To answer our research questions, we conducted semi-structured interviews. The participants need to own or have consistent access to PI systems, such as Fitbit, iWatch, Garmin, Strava, etc., for at least three months and have reviewed their systems' dashboard(s) and report(s) for self-reflection at least once a month. Twenty individuals participated in the study. In addition, a brief survey collected demographic information and had users rate (Likert scale) the impact of missing data on their PI goals and overall attitude towards missing PI data.

3.2 Participants

We sent our recruitment email to private training groups, Panther Cycling Club and Club Triathlon of Pittsburgh. We also sent out the recruitment questionnaire to a crowdsourcing website, Prolific, to recruit participants who have been using PI tools. As a result, 314 people responded to the questionnaire. We applied the inclusion criteria, which are 1) used PI tools for more than three months, 2) reflected on their data at least once a month, 3) collected data towards a specific goal, and randomly selected participants across the three recruitment sources. As a result, 20 people (11 female, 9 male) participated in the study, and the participants were aged 19 - 56 years, M = 32.875, SD = 9.34, and resided in the US. The number of participants was informed by the Qualitative Research and Saturation Criteria [30]; we concluded the surveys when no new information emerged from later interviews. As represented in Table 1, all participants had relatively extensive tracking experience.

3.3 Procedure

The design of this study was approved by our University's institutional review board (IRB). The potential participants who took the pre-screening Qualtrics survey received an email to schedule a virtual interview for up to an hour. This work was conducted during the COVID-19 pandemic; interviews were conducted over Zoom. We recorded the audio of the interviews. In the interview, we first gained informed consent and then engaged in a series of questions to have the participants describe the PI tools they used, their motivation, and how they used them. We then asked them to show us how they reflected on their data and how missing data would occur and influence this process. Then, we asked them to express their attitudes toward missing data in descriptive sentences and on a 5-point Likert scale. Finally, we asked them for suggested methods to mediate missing data. We compensated the participants $15 for their participation.

3.4 Data Preparation and Analysis

The audio files from the interview were imported and machine transcribed using the cloud-based platform Atlas.ti. Two researchers (both authors on the paper) used an open coding technique to identify themes and trends in the responses, and proceeded to identify relationships among the codes (axial coding [32]). Then the first researcher and a third researcher (not an author on the paper) separately read and coded 10% of the transcripts using the themes identified previously. Their initial inter-rater reliability (percentage agreement) was 0.769; which is higher than the expected agreement by mere chance, 0.56, proposed by Krippendorff ([15], p. 224–226). The first and the third researcher discussed discrepancies and updated the existing codebook. Afterwards, the first and a fourth researcher (not an author on the paper) coded the rest of the transcripts, and the inter-rater reliability by Krippendorff's alpha was 0.73, and the agreement was 0.99.

4 Results

Our study found notable user behaviors and perceptions in situations where data are missing in personal informatics tools. Specifically, several themes emerged that link missing data to capture and goal-tracking challenges. Within these themes, we observed two very distinct classes of users: ***trainees*** and ***maintainers***. Trainees are individuals with concrete, even professional athletic training goals who seek data to guide decisions to improve performance-related metrics (e.g., cardiac efficiency, time on interval). Maintainers have broad health improvement and maintenance goals. They seek data to track milestones (e.g., steps per day, active exercise hours). Given these very divergent goals, we present results from the perspective of *both* classes of users.

4.1 Usage of the PI Tools

To provide context, we first describe how trainees and maintainers used PI tools as part of their daily routines. A wide variety of PI tools were used across the 20 participants interviewed. Table 2 summarizes these tools, separating them between the two classes of users. We performed an analysis of each tool, assessing the presence of functionality along four dimensions: 1) data export, e.g., use of

Table 1. Participants demographics, tracking background, and frequency of reflection

PID	Age range	Gender	Occupation	Wearables	Duration	Category	Main motive of use	Frequency and types of reflection
P1	32–43	Male	Cycling coach	Garmin watch	13 years	Trainees	Training performance, analysis for cycling data, balance training load.	Reflect on physiological data after biking, deeper analysis on the weekend to check performance, modify training load, and analyze performance per season and year
P2	19–31	Female	Service	Apple watch	7 years	Maintainers	Maintain heart rate, maintaining weight, be active.	Check data after workouts, check meal data daily on the app for nutrition, reflect on weekends for meal summary
P3	32–43	Male	Teacher	Garmin Phoenix five	13 years	Trainees	Replicate best performance for the race, inform training schedule	Reflect on data after running, reflect on weekly data to analyze and identify best performance based on road conditions, pace, and duration
P4	32–43	Male	Software engineer	Kronos watch	1 year	Maintainers	Maintaining weight, improve health	Check to see if hit daily targets, check weekly data to see trends
P5	44–56	Female	Unemployee	VeryFit watch	2 years	Maintainers	Maintaining health, losing weight	Reflect on weekly data and daily data to see if hit targets
P6	19–31	Female	Desk job	Apple watch	> 6 years	Maintainers	Tracking activities, checking calories burned.	Multiple times a day to see if they hit daily targets, reflect three times a month to see summaries reflect, weekly to see if they hit weekly targets
P7	19–31	Male	Student	Suunto watch	3 years	Trainees	Tracking and analyzing cycling data, log exercises.	Multiple times a day, reflect after cycling for performance
P8	19–31	Female	Student	Garmin watch	2 years six mo	Maintainers	Tracking activities, hit exercise target.	Check during workouts, reflect at least once a day to see if hit targets
P9	32–43	Female	Unemployee	Fitbit	> 10 years	Maintainers	Tracking activities, log workouts.	Check data after workouts, check weekly to see exercise types and duration
P10	44–56	Male	Learning Consultant	Garmin watch	5 years	Trainees	Maintain certain pace during marathon	Reflect after running for pace, heart rate and road condition
P11	19–31	Female	Desk job	Garmin watch	5 years	Trainees	Training to advance calisthenics	Reflect three times a week for how many times reached the goal
P12	32–43	Female	Desk Job	Fitbit	5 years	Maintainers	Tracking exercises, maintaining weight, being active	Check multiple times a day during exercise, reflect once a day to see if hit targets
P13	19–31	Female	Unemployee	Fitbit	1 year	Maintainers	Tracking activities, be active	Checking data multiple times a day and during exercise, reflect daily to see if hit targets
P14	19–31	Female	Student	Apple Watch, Whoop band	4 years	Trainees	Training for Brazilian Jiu-Jitsu competition, improve performance	Reflect daily to check recovery score, inform the types of training the body is ready for the day, reflect weekly and monthly to check trends/patterns of the week&month
P15	19–31	Female	Student	Apple Watch	2.5 years	Maintainers	Tracking exercises, losing weight	Reflect daily to check if hit targets, check during exercise to see progress
P16	19–31	Female	Teacher	Apple Watch	2 years	Trainees	Training for weight lifting, building muscles	Checking after training for performance: duration, weight; reflect to inform training load
P17	32–43	Male	Pilot	Whoop band	1.5 years	Trainees	Training for hike Mount Rainier	Reflect after hiking for speed, heart rate, and altitude of the mountain
P18	44–56	Male	Maintenance technician	Garmin Vivoactive Three	8 years	Trainees	Training for running	Checking data for time and heart rate during training. Balance training load. Reflect after running for hill work, pace, distance, elevation, and duration
P19	32–43	Male	Teacher	Misfit Vapor X	4 years	Trainees	Training for cycling, making sure to hit endurance targets	Reflect weekly to inform training in the following week, reflect after cycling to check for performance
P20	19–31	Male	Wine tasting host	Apple Watch	> 1 year	Maintainers	Tracking workout, stay active	Checking data multiple times per day for progress, reflect weekly for summaries

data captured by the tool for use in another tool, 2) data visualization, e.g., visual depictions of activity and progress towards a defined goal, 3) data analysis, e.g., identify performance trends and calibrate future goals, and 4) social features, e.g., posting on social media a physical activity or accomplishment.

Table 2. PI wearables/applications used by participants. "●" means feature was being used by at least one participant, "○" means features was not being used by participants who reported using the wearables/applications, "—" means device does not have such a feature

Category	**PI tools used**	**Usage of the PI tools**			
		Data export	Data visualization	Data analysis	Social features
Trainees	Strava	●	●	●	●
	Pedometer	●	—	—	—
	Google Fit	●	○	○	—
	Power meter	●	—	—	—
	Whoop band	●	●	●	○
	Couch to 5k	●	—	—	—
	Suunto watch	●	○	●	○
	Apple watch	●	○	●	○
	Garmin watch	●	○	●	○
	Golden cheetah	—	●	●	—
	Bike computers	●	○	—	—
	GPS foot pedals	●	—	—	—
	Garmin Fenix watch	●	○	●	○
	Garmin Vivoactive watch	●	●	●	○
Maintainers	Lose it	○	●	○	—
	Virgin Pulse	○	●	○	—
	Apple watch	●	●	●	●
	Fitbit watch	●	●	○	●
	Veryfit watch	○	●	○	—
	Kronos watch	●	●	●	●
	My fitness pal	○	●	○	○
	My Maintainers Pal	○	●	○	○
	Apple health	○	●	○	●
	Garmin Connect	○	●	○	●
	Garmin Vivoactive watch	○	●	○	●

The analysis of feature presence included reflection on participants' explicit descriptions of use, our assessment of manufacturer marketing materials and technical manuals, and (if application-based) our independent installation and exploration. The four latter columns of Table 2 report the result of this analysis. A circle indicates we assessed the feature dimension to be present in the tool, and a dash, if assessed, is not present. A solid shading indicates at least one participant indicated using the feature. An empty circle indicated the feature was present, but observed no reported use among participants in the interviews.

We discovered several interesting patterns through this analysis. Most prominent, many of the *same* tools were used in different ways by the different classes of users. Specifically, ***trainees*** overwhelmingly favored the use of data export and data analysis functionalities over the built-in (or even *non-present*) data visualization features. The opposite use pattern was observed for ***maintainers***, who favored visualization features. We captured sentiment and motivation for these divergent behaviors in the interviews.

Interviews explored the lifecycle of personal informatics tools. As expected, tools are most commonly discontinued when users acquire new, more capable tools (e.g., upgrading to the latest fitness watch), resulting in missing data in their records.

Table 3. Reflective behaviors

Category	**Motive of use**	**Data Usage (goal)**	**Frequency of use**		**Features used**
			Collection	Reflection	
Trainees	Cycling Marathon Martial arts Coach others Weight lifting Cycling tournament Running tournament	Training power Training strength Training sprint Analyze performance Balance training load Inform training plan Optimize performance Improve performance Maximize abilities in running Replicate best performance	Upon training	Daily, upon training, weekly, monthly	Identify important moments; Training stress balance; Chronic training load; Elevation gain; Speed; Power; Heart rate zones; Cadence; Pace; Training intervals; Identify min/max value for each feature; Recovery score.
Maintainers	Educate self Log activities Improve health Maintain health Manage weight Maintain heart health	Map routes Plan meal Monitor Sleep Lose weight Maintain weight Check Mileage Regulate heart rate Check step counts Check Calories burned Monitor Sleep conditions Log Swimming hours	Daily	Weekly, monthly	Calories intake and nutrition; protein and carbohydrates in each meal; Active hours; Sleep duration per day; Total workout hours; Check completion rate; Check targets set for activities; Compare performance.

4.2 Users' Expectations of Consistency and Completeness in PI Data

Knowing how people utilize their PI tools for different motivations (see Sect. 4.1, Table 3 further discriminates participants' usage behaviors into ***trainees*** and ***maintainers***. We analyzed the features and frequency of use, during two stages: 1) collection, e.g., collecting sensory data using the PI tools, and 2) reflection, e.g., reflecting on the collected data to gain actionable insights for behavior change.

The analysis of the frequency and features used during reflection is extracted from participants' explicit descriptions of how they reflect. The last two columns report the result of this analysis. A notable distinction emerged between the different classes of users. ***Maintainers*** would regularly use and interact with the tools at the time of collection, often to confirm data collection and assess progress towards the goal. In contrast, the ***trainees*** were found less engaged, often checking the tools only once a day or periodically throughout the week to reflect on their activity trends and broader health maintenance goals.

Reflection behaviors with ***trainees*** often centered on *understanding the past* and *predicting the future*. For instance, P1, a trainee, stated the need to guide adjustments in a training routine, by reflecting on previous performance data: *"if somebody is losing races and their sprint is not as good as before, he/she will need more sprint work (recommended method of choice for cardiovascular exercise)"*. P3, also a trainee, noted he believes *"it helps you predict how fast you can actually race because you know exactly how long you can hold at a certain* *heart rate for"*. P13, a trainee again, stated the need to maintain a reasonable training stress balance: *"based on how recovered I am, this is how much strain I should put on my body"*.

In comparison, ***maintainers*** would often reflect to assess past behaviors against goals *knowing the present*. For example, P6, a maintainer, commented *"usually just checking how far I am from the target"*. In addition, 7 out of 10 maintainers *(P5, P6, P9, P4, P13, P8, and P20)* stated they mainly log how many exercises, or how many times for an exercise session, to see if they met the perfect week (hit targets every day), the perfect span, exercise, and weekly goals.

The value of precision and detail in PI also differed. ***Trainees*** were found to be interested in specific insights and would compare exact data across multiple periods or durations (e.g., daily, weekly, and monthly). For example, trainees use important moments, chronic training load, training intervals, and recovery scores to balance easy and heavy training. P3 noted that knowing the min/max of all these features helps him *"analyze from an analytical perspective about what my body is capable of"*, and claims one can even replicate a certain level of heart rate and pace to mimic past victories in a new tournament. On the other hand, ***maintainers*** focused on summative insights across the broad set of past activities. As the last column shows, features like active hours, sleep duration, total workout hours, etc., are oriented toward tracking trends and broad health goals.

4.3 Missing PI Data Conflict and Users' Expectations

The reflective behavior described in Sect. 4.2 is dependent on the PI data collected. We noted the presence of missing data in either of the stages to be problematic. For ***trainees***, activities that had no data prevented effective comparisons across activities or tracking specific performance metrics. For ***maintainers***, a lack of data could incorrectly imply long-term goals are not being achieved. Across both groups, the absence of key data was linked to the abandonment of the devices, indicating little tolerance for adapting to the functional limitations of the tools.

Both ***maintainers*** and ***trainees*** claimed they had experienced missing data when using the PI tools in our study. We analyzed the causal factors that lead to missing data and its effects during the lifecycle of personal informatics. Our analysis revealed two dimensions: 1) human reasons, e.g., missing data caused by participant error or behavior; 2) device reasons, e.g., missing data caused by malfunction or misconfiguration of wearables or applications. These dimensions apply in both groups.

Missing data caused by humans are: 1) *Forgetting to initiate data collection*, e.g., participants would forget to invoke the application or device to collect data for an activity. This could also include forgetting to annotate data or input manual entries. 2) *Forgetting to bring the device*, e.g., participants would forget to bring or wear the device for an activity. In some cases, like for P4, a ***maintainer***, and P8, a ***maintainer*** again, who forgot to bring their smartwatch on vacation,

Table 4. Quantifiable questions

Questions	Ratings
Q1. Please rate your frustration level when the data is missing	“Not at all”
Q2. Please rate your frustration level when the data is inaccurate	“Slightly”
Q3. Please rate your trust level towards your tracking data.	“Somewhat”
Q4. Please rate the level of influence missing data has on your goal.	“Moderately”
	“Extremely”

these gaps can be over many days. Most common, was forgetting to put the device back on after charging.

Missing data caused by devices include: 1) *Battery died*, a significant portion of the participants experienced battery issues during their activities. e.g., P8, ***maintainer***, claims *“sometimes I forget to charge my watch, it’ll die in the middle of a workout or run”*. 2) *Malfunction or limitation of the device*, e.g., P6, a ***maintainer***, reported that her smartwatch did not count the swim laps. 3) *Syncing problem*, e.g., P7, a ***trainee***, stated that he uses multiple applications, and when syncing the data to other applications, data points were lost. 4) *Precision of the device*, all participants claimed that their device would often not capture activities at the correct level of precision, e.g., P4, a ***maintainer***, claimed that sometimes, when he was holding the wheel and driving, the smartwatch would pick up the vibration and count it as steps. P16, a ***trainee***, stated *“sometimes, if I’m walking around the class and talking really loud, my heart rate might spike a little bit, but it (smartwatch) may count that I’m actively working out, which I’m not”*.

The quantitative analysis on the 5-point Likert scale rating (questions shown in Table 4) showed that 20% maintainers and 40% trainees claimed missing data does not affect their goals (Q4, rating of “Not at all”), 60% maintainers and 40% trainees reported that missing data slightly influences their goals (Q4, rating of “Slightly”), 20% maintainers claimed missing data somewhat affects their goals (Q4, rating of “Somewhat”), and 20% maintainers reported missing data moderately or extremely influence their goals (Q4, rating of “Moderately”).

20% maintainers and 30% trainees reported taking action when they noticed the missing data, and the same percentage of participants also claimed that if the data is missing, they will refer to their friends’ data who were on the same route to estimate their PI data. Furthermore, 20% of the trainees would estimate the data based on how their body feels. In the reflection stage, if participants noticed the tools did not capture their hours/activities during the day, they would redo the workout to make up for the missed data. For those estimated data, they reflect on it with no difference from regular data, but some participants also claimed that they prefer no data to inaccurate data under some circumstances, e.g., P18, a ***trainee***, claimed *“The estimated calories burned are accurate to about 90%, which is fine. But for the estimated heart rate? No, absolutely not. I would rather the app (Google Fit) tell me it did not capture it (heart rate). ”*

Interestingly, having data towards a goal would even impact some participants to repeat activities to properly capture the PI data. For instance, P8, a ***maintainer***, who claimed missing data somewhat affects her goal, commented that *"If I set 100 (minutes), and I had missing data and the watch shows I would have gotten 30 (minutes). I'll probably just run again with my watch, and then get more minutes, just so I can reach my goal (set) on the app"*.

Frustration with missing data (and, by consequence, inaccurate data) was also varied, 10% trainees and 20% maintainers claimed they felt somewhat frustrated when their data was missing (Q1, rating of "Somewhat"), 30% trainees and 20% maintainers stated their frustration towards missing data is moderate (Q1, rating of "Moderately"), and 30% trainees rated their frustration as extremely (Q1, rating of "Extremely"), an equal percentage of 20% in both groups expressed as slightly (Q1, rating of "Slightly"), and 10% trainees and 30% maintainers said they were not at all frustrated about missing data (Q1, rating of "Not at all"). All participants acknowledged that the data collected had some discrepancies, and it is not technically practical to have data with 100% accuracy. P5, a ***maintainers***, commented *"I like things to be accurate, I won't say I'm a number person per se, but I like things organized and proficient, so it (inaccurate data) did bother me at first"*. While acknowledging PI data is not 100% accurate, overall levels of trust in collected PI data were high. 50% trainees and 30% maintainers trusted their data extremely (Q3, rating of "Extremely"), 30% trainees and 50% maintainers trusted moderately (Q3, rating of "Moderately"), and 20% trainees and maintainers in each group trusted the data at a somewhat level (Q3, rating of "Somewhat").

From this analysis, we noticed that although missing data impact participants negatively in both groups, there are minor differences regarding the level of effects. For example, participants in the trainees' group tend to be more frustrated when their data is missing. On the other hand, the maintainers' group is more frustrated if the data is inaccurate.

5 Implications

From the semi-structured interview, we found different usage patterns for trainees and maintainers; we identified that missing data causes conflicts. We also elaborated on how their expectations were different between these two groups.

Our results contextualized the impact forgetting to wear or charge devices has on the value and utility of PI tools. This is a *fundamental* design limitation of PI technology – there is no clear technology horizon where these tools and devices automatically charge and attach to users. Designing *for* these events is critical to the long-term use and the utility of PI. Thus, in this section, we address the question: "how can the design of PI tools be extended to help users reflect upon and mitigate missing data in their tracking activities?"

Here we propose to take synthetic data as a principle when designing PI tools. By synthetic data, we meant to provide a visual representation in helping users to distinguish missing data vs. no data collected. And assist users in estimating their missing data during the reflection process.

Our results pointed to several key implications on how PI tools could be improved to embrace synthetic data in the user experience. Including synthetic data (missing data estimation) in the design principle of PI means the tool should not only consider where to include it in the tool, how to implement it based on different motivations, but also on how to represent it in the tool.

5.1 Usage Behavior

Our study showed three usage behaviors: 1) *understanding the past*, 2) *knowing the present*, and 3) *predicting the future* (see Sect. 4.2). Usage behaviors towards collected data are essential to self-reflection and finding actionable insights for behavior change. Our analysis indicated that current tools lack support for integrating that three usage behaviors. As a result, participants use separate PI tools for different use cases. This implies tools that enable synthetic data should consistently represent: 1) where synthetic data exists of weekly, monthly, and yearly data; 2) how synthetic data is used to estimate goal tracking; 3) control the use of synthetic data to discover patterns used to inform future behavior towards specific events or goals. All three usage behaviors are interconnected; a consistent representation of how missing data is represented and interpreted can reduce fallacious insight and increase users' confidence in their data.

5.2 Sensitivity to Missing Data

The result showed diverse levels of tolerance to missing data by participants (see Sect. 4.3). For example, some participants claimed they could accept one or two days of missing data per week. Some tolerate only a few hours per day. Participants consistently expressed tolerance for missing data within the context of interference with tracking goals, trends, and specific activities. This finding suggests no easy solution to meet missing data sensitivity through a single mitigation approach. Feedback from participants suggests three design approaches: 1) Allow the user to fill in the missing data manually; ideal for small gaps where a specific activity was performed (e.g., an outdoor run). 2) Use algorithmic mediation to estimate gaps in data based on prior data collection; ideal for large data gaps in daily, repeated activities (e.g., determining steps taken on a routine evening walk). 3) Use algorithmic mediation and guiding user interfaces to help users refine estimates based on known conditions of the missing data; ideal for larger gaps where a specific activity or set of activities was performed) (e.g., determine the performance of a long-distance run based on knowing the start and stop times and locations). Future work is needed to understand the effectiveness of each approach, and under what conditions each approach is preferred by users.

5.3 Visual Representation of Missing Data

For all the PI tools reported in use by participants in our study (See Table 2), the analysis indicated the tools handle missing data as the absence of data.

Exercise or activity could have been performed, but because data on the event was not collected, the tools often convey to the user *no activity* was conducted. This is misleading and can incorrectly impact a user's personal health goals. Tools should provide a better visual distinction between when data is missing and when activity is simply low. Identifying gaps when a device was worn can help users contextualize the data in the view, allowing them to reflect better on their physical activities. Further, visualizing gaps can be a natural opportunity in the user experience to allow users to estimate missing data manually or to enable an algorithm to generate heuristically derived synthetic data.

5.4 Trust in Data

Our analysis showed that participants largely believe the PI data collected is accurate (see Sect. 4.3). Trust towards the PI tools is critical to adherence and continued use. Participants would identify other sources for direct or interpretive comparison to further verify this trust. For instance, participants would compare data of individuals that also performed the same activity, examine the map while they ran or walked on a specific route in real-time, and estimate data reflecting on the tiredness of their body. These behaviors imply tools that enable synthetic data creation should enable users to: 1) observe synthetic data within the context of plausible and prior representative behaviors to assure synthetic data is grounded in primary data (e.g., comparison to previous five similar activities); 2) group and label synthetic data as specific, personally relevant events (e.g., labeling 'run with John'); and 3) be able to manipulate and revise algorithmic estimates based on intuition or preference in how missing data should be created.

6 Conclusion

This work considered a new perspective (missing data) on personal informatics and investigated how missing data impacts user behaviors and perceptions in authentic, everyday use. The work extended the community's understanding of the circumstances that cause missing data and how those circumstances influence collection and reflection behaviors. Our analysis demonstrated the issues caused by missing data at the level of individuals and technical issues. In addition, it contributed new insights into how personal informatics tools should adapt to support synthetic data and provided insights to guide how they can do it.

During the analysis, we identified two distinct groups, trainees and maintainers. We found trainees mostly use their PI tools to *understand their past* and *predict the future*. In contrast, the maintainers often use their PI tools to *know the present*. For both groups, we have provided a comparative result on missing data's influence on the lifecycle of personal informatics, and the implications capture lessons from the perspective of both groups. The resulting analysis point to key limitations of current tools, which is the lack of representation for missing data, and outlined design guidelines for future tools to improve the user experience with PI data. For future work, we would need to implement different methodologies based on the implementations to generate and represent synthetic data, conduct user studies to investigate the effectiveness of each approach, and determine under what conditions each approach is preferred.

References

1. Agapito, G., et al.: DIETOS: a recommender system for adaptive diet monitoring and personalized food suggestion. In: 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 1–8. IEEE (2016)
2. Alvarez-Lozano, J., et al.: Tell me your apps and i will tell you your mood: correlation of apps usage with bipolar disorder state. In: Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments, pp. 1–7 (2014)
3. Baumer, E.P., Khovanskaya, V., Matthews, M., Reynolds, L., Sosik, V.S., Gay, G.: Reviewing reflection: on the use of reflection in interactive system design. In: Proceedings of the 2014 Conference on Designing Interactive Systems, pp. 93–102 (2014)
4. Choe, E.K., Lee, B., Kay, M., Pratt, W., Kientz, J.A.: SleepTight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 121–132 (2015)
5. Choe, E.K., Lee, B., Zhu, H., Riche, N.H., Baur, D.: Understanding self-reflection: how people reflect on personal data through visual data exploration. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, pp. 173–182 (2017)
6. Choe, E.K., Lee, N.B., Lee, B., Pratt, W., Kientz, J.A.: Understanding quantified-selfers' practices in collecting and exploring personal data. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1143–1152 (2014)
7. Choi, W., Park, S., Kim, D., Lim, Y.K., Lee, U.: Multi-stage receptivity model for mobile just-in-time health intervention. Proc. ACM Interact. Mobile, Wearable Ubiquit. Technol. **3**(2), 1–26 (2019)
8. Chung, C.F., et al.: Boundary negotiating artifacts in personal informatics: patient-provider collaboration with patient-generated data. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pp. 770–786 (2016)
9. Consolvo, S., et al.: Flowers or a robot army? Encouraging awareness & activity with personal, mobile displays. In: Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 54–63 (2008)
10. Consolvo, S., McDonald, D.W., Landay, J.A.: Theory-driven design strategies for technologies that support behavior change in everyday life. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 405–414 (2009)
11. Epstein, D.A., Jacobson, B.H., Bales, E., McDonald, D.W., Munson, S.A.: From " nobody cares" to "way to go!" a design framework for social sharing in personal informatics. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 1622–1636 (2015)
12. Epstein, D.A., et al.: Examining menstrual tracking to inform the design of personal informatics tools. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 6876–6888 (2017)

13. Epstein, D.A., Ping, A., Fogarty, J., Munson, S.A.: A lived informatics model of personal informatics. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 731–742 (2015)
14. Fox, S., Howell, N., Wong, R., Spektor, F.: Vivewell: speculating near-future menstrual tracking through current data practices. In: Proceedings of the 2019 on Designing Interactive Systems Conference, pp. 541–552 (2019)
15. Friese, S.: ATLAS. ti 8 windows-inter-coder agreement analysis copyright 2019 by atlas. ti scientific software development GmBH, berlin. all rights reserved. manual version: 652.20190425. Updated for program version: 8.4 (2019)
16. Fritz, T., Huang, E.M., Murphy, G.C., Zimmermann, T.: Persuasive technology in the real world: a study of long-term use of activity sensing devices for fitness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 487–496 (2014)
17. Gouveia, R., Karapanos, E., Hassenzahl, M.: How do we engage with activity trackers? A longitudinal study of habito. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1305–1316 (2015)
18. Hsieh, C.K., et al.: Lifestreams: a modular sense-making toolset for identifying important patterns from everyday life. In: Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems, pp. 1–13 (2013)
19. Lazar, A., Koehler, C., Tanenbaum, T.J., Nguyen, D.H.: Why we use and abandon smart devices. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 635–646 (2015)
20. Li, I., Dey, A., Forlizzi, J.: A stage-based model of personal informatics systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 557–566 (2010)
21. Li, I., Dey, A.K., Forlizzi, J.: Understanding my data, myself: supporting self-reflection with ubicomp technologies. In: Proceedings of the 13th International Conference on Ubiquitous Computing, pp. 405–414 (2011)
22. Lin, J.J., Mamykina, L., Lindtner, S., Delajoux, G., Strub, H.B.: Fish'n'Steps: encouraging physical activity with an interactive computer game. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 261–278. Springer, Heidelberg (2006). https://doi.org/10.1007/11853565_16
23. Martin-Niedecken, A.L., Rogers, K., Turmo Vidal, L., Mekler, E.D., Márquez Segura, E.: ExerCube vs. personal trainer: evaluating a holistic, immersive, and adaptive fitness game setup. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–15 (2019)
24. Miller, A.D., Mynatt, E.D.: StepStream: a school-based pervasive social fitness system for everyday adolescent health. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2823–2832 (2014)
25. Min, J.K., Doryab, A., Wiese, J., Amini, S., Zimmerman, J., Hong, J.I.: Toss'n'turn: smartphone as sleep and sleep quality detector. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 477–486 (2014)
26. Mishra, V., Künzler, F., Kramer, J.N., Fleisch, E., Kowatsch, T., Kotz, D.: Detecting receptivity for mHealth interventions in the natural environment. Proc. ACM Interact. Mobile, Wearable Ubiquit. Technol. **5**(2), 1–24 (2021)
27. Rapp, A., Cena, F.: Personal informatics for everyday life: how users without prior self-tracking experience engage with personal data. Int. J. Hum Comput Stud. **94**, 1–17 (2016)

28. Rooksby, J., Rost, M., Morrison, A., Chalmers, M.: Personal tracking as lived informatics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1163–1172 (2014)
29. Saksono, H., Castaneda-Sceppa, C., Hoffman, J.A., Seif El-Nasr, M., Parker, A.: StoryMap: using social modeling and self-modeling to support physical activity among families of low-SES backgrounds. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2021)
30. Saunders, B., et al.: Saturation in qualitative research: exploring its conceptualization and operationalization. Qual. Quant. **52**(4), 1893–1907 (2018)
31. Schon, D.A.: The Reflective Practitioner: How Professionals Think in Action. Basic books, New York (1983)
32. Scott, C., Medaugh, M.: Axial coding. Int. Encycl. Commun. Res. Meth. **10**, 9781118901731 (2017)
33. Sellak, H., Grobler, M.: MHealth4U: designing for health and wellbeing self-management. In: 2020 35th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW), pp. 41–46. IEEE (2020)
34. Sharmin, M., et al.: Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 505–516 (2015)
35. Theodoridis, T., Solachidis, V., Dimitropoulos, K., Gymnopoulos, L., Daras, P.: A survey on AI nutrition recommender systems. In: Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments, pp. 540–546 (2019)
36. Franco, R.Z.: Online recommender system for personalized nutrition advice. In: Proceedings of the Eleventh ACM Conference on Recommender Systems, pp. 411–415 (2017)

"I Have to Do Something About It" - An Exploration of How Dashboards Invoke Self-Reflections in Chronic Obstructive Pulmonary Disease Patients

Stephanie Githa Nadarajah, Peder Walz Pedersen, Milo M. Skovfoged, Hamzah Ziadeh, and Hendrik Knoche(✉)

Department of Architecture, Design, and Media Technology, Aalborg University, Aalborg, Denmark
`hk@create.aau.dk`

Abstract. Chronic Obstructive Pulmonary Disease (COPD) patients need to track their symptoms for health professionals to adapt treatments in a timely manner in case of health deterioration. Clinicians typically analyzed the tracked data and recommended actions to patients who acted as mere data collectors. Consequently, patients have little agency and motivation to self-track. Two studies investigated how digital dashboards influenced patients' motivation, agency, and reflections. Study 1 (one week) focused on how five patients used a paper diary to self-track and reflect on their symptoms. Additionally, the patients evaluated a tablet-based digital dashboard using four data visualisations. Study 2 looked at how five patients tracked and reflected on their data using a tablet-based dashboard for two weeks. By using reflective questions to prompt patients to compare and reflect on time series charts with data annotations, patients gained new knowledge about what factors might influence their symptoms and identified actions to improve their health (e.g. increase oxygen supplements). This strengthened their sense of agency and motivated them to participate more in the management of their condition.

Keywords: Self-tracking · chronic obstructive pulmonary disease · COPD · personal informatics · data collection · reflection · dashboard

1 Introduction

To prevent the health decline of patients with chronic obstructive pulmonary disease (COPD), healthcare professionals need to monitor their condition and symptoms. Some countries (e.g. Denmark) and hospitals require patients to submit data on symptoms, so-called patient-reported outcomes (PRO), for clinician analysis and decision making on when to take actions [29,32,40]. This gives patients little agency or insight into their health, which reduces their motivation to collect health data consistently [17,25]. Patients also lacked knowledge on how to interpret and reflect on health data to

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 619–635, 2023.
https://doi.org/10.1007/978-3-031-34586-9_41

improve their condition which impeded their self-tracking [17,25,26,29]. Some studies hypothesized that providing instructions on self-tracking to encourage patients' reflection can increase their motivation to self-track and enable them to improve their health [25,26,29]. Current digital dashboards have employed visualisations, written guides, and reflective questions to instruct patients on how to collect and reflect on data [1,20,29]. While many studies investigated how patients interpreted dashboards in single sessions lasting no longer than one hour [3,24,29,32], only few investigated how reflecting on self-tracked data can impact COPD patients' motivation and knowledge over a longer period of one or more weeks [25,26].

This paper describes two studies investigating how a digital dashboard for COPD patients impacted knowledge and motivation for self-tracking after extended use. Our findings suggest that patients can gain new knowledge about how to track and alleviate their symptoms after using a dashboard promoting reflections for two weeks. We describe how reflections self-tracking gave COPD patients a sense of agency over their illness and motivated them to actively reflect on and improve their health.

2 Background

Healthcare professionals (i.e., clinicians) cannot cure or reverse COPD but can administer treatments to reduce symptoms. Clinicians relied on PRO measures through telehealth applications to initiate these treatments and prevent health declines [32,40]. In the absence of a model for such mandated use, we relied on Li's stage-based model of personal informatics systems [20] to define the activities of self-tracking. The model breaks down self-tracking into five stages: 1) determining variables, tools, and frequency of tracking (*preparation*), 2) logging data (*collection*), 3) preparing data for reflection e.g. by aggregating and analysing data (*integration*), 4) examining data to generate knowledge (*reflection*), and 5) deciding what actions to take (*action*) [20].

In the COPD case, clinicians prepared the collection stage by predefining relevant symptoms for tracking through objective numerical measures (e.g. oxygen saturation measures) and subjective binary measurements (e.g. yes/no answers to whether dyspnea has increased more than usual). Clinicians integrated and reflected on the patient-provided data to determine whether these fall into acceptable ranges and advise patients on possible actions [29,32,40]. Thus, patients only received feedback when clinicians identified declining health which demotivated some patients to incorporate self-tracking into their daily habit as they could not reflect on their data themselves [1,17,41]. To gain a sense of agency, patients recorded data for their own use using notebooks, diaries, and applications [16,27–29,31]. However, many patients lacked the knowledge and skills to reliably track data and identify variables impacting their results (e.g. weather) [1,7,11,30,40]. Both healthy users [2,13,36] and patients [17,25,29] did not know how to reflect on their data nor identify appropriate actions to improve their health (e.g. losing weight). Instead, self-trackers; including patients, needed actionable (expert) advice [20,30,41].

Self-trackers often lost their motivation due to tracking fatigue caused by the continuous effort needed to measure and log data [6,20,28]. Chronically ill patients (e.g. with cancer) experienced fatigue more strongly due to their symptoms, prompting them to stop self-tracking earlier than healthy users [1,31]. Demotivation may lead to patients postponing the recording data [1,23], which biased later entries and reduced data reliability [18]. Alternatively, patients measured symptoms in a disorganised manner resulting in poor self-tracking data (e.g. incorrect measures) [3,19,40]. Both healthy [11,20] and COPD [17] self-trackers felt motivated by monitoring progress towards goals and their curiosity in the data which lead to concrete actions towards health improvement.

These actions stemmed from either short-term reflections where self-trackers reflected on their status immediately after data entry or long-term reflection where self-trackers reflected over trends after several days or weeks [20]. Long-term reflection allowed for higher levels of data exploration by comparing, exploring, and finding patterns [20]. Typically, reflections arose from discrepancies between actual and expected measurements (e.g. recommended levels, goals, and previous levels) [1,13,27,35]. To trigger reflections, some systems guided patients' towards discrepancies in the data through reflective questions asking patients to explain discrepancies [17,27,29]. Presenting data through data visualisations aided healthy users in exploring patterns to discover discrepancies [6,13,20]. Therefore, self-tracking applications employed data visualisations to aid both healthy and chronically ill users in short-term reflections about their health [5,12,34,44]. However, data visualisations must account for the users' condition, skill, purpose, and motivation to aid reflection [9,28]. For example, data visualisation can support the questions self-trackers pose while reflecting: 1) What is my current status? 2) How does the current status compare to earlier values (history)? 3) How is it related to other variables? 4) What affects my current status? 5) What is an appropriate goal? 6) How does the current status compare to my goal? [21].

Single value charts provided self-trackers with a quick overview of their current status [28]. Time series charts visualized past experiences (history) to reflect on trends or deviations and triggered storytelling about experiences behind data [1,28,35]. Comparison charts with time series visualisations for multiple variables sharing the same vertical axis supported reflection on discrepancies between variables [9,38]. Calendar heatmap visualisations can illustrate periodic patterns by color-coding variables [9]. Baselines (e.g. general averages) added to visualisations provided context to measurements when reflecting [15] and helped chronic patients determine the severity of their symptoms, which they previously found difficult as they were constantly symptomatic [17,40]. However, self-tracking applications often limited reflections by only having either simple visualisations invoking short-term reflection or more complex charts (time series) invoking long-term reflection [21,23].

Dashboards can contain multiple simple and complex visual visualisations of the self-tracked data [9]. Most dashboards only allowed for explanatory analysis through static charts which they could not create, search, or edit [6,7,20,42]. While this limited users' knowledge generation, it simplified the task of reflecting for users unfamiliar with visualising and analysing data [37]. Both clinicians and patients valued dashboards for patients to reflect on their condition [15,29,39]. Despite this, only a few studies investigated how dashboards affected patients' self-tracking [22,24]. Patients often struggled to generate knowledge from dashboards as they did not understand the medical terminology and data [22,24]. When confronted with visualisations depicting worsening health (e.g. increased symptoms, decline in physical or cognitive performance, etc.), patients tended to reject results or stopped reflecting [1,15,17,29]. However, the majority of these studies investigated self-tracking contained in single sessions, typically less than an hour [15,22,24]. They focused on short-term use and did not investigate how patients adapted to tracking, knowledge development over time, and motivating factors of continued self-tracking [26].

This paper investigates how dashboards utilizing data visualisations, contextual annotations, and reflective questions can support COPD patients to reflect on their condition, thereby improving their sense of agency and motivation to self-track.

3 Study 1 - Reflecting on a COPD Dashboard

This study explored COPD patients' initial opinions of a web-based dashboard designed to encourage reflections. To track their health for a week, participants received a paper diary that utilized different ways to prompt reflection. Afterwards, participants reviewed four dashboards designed to promote short and long-term reflections using different, simple and complex data visualisations paired with reflective questions and contextual annotations to prompt reflection. We evaluated, which designs patients preferred through in-situ interviews.

4 Apparatus

The diary (see Fig. 1) consisted of three assignments conducted over a week (Monday, Wednesday, and Friday). On each day, participants recorded their: pulse, weight, blood oxygen level, and answered two questions: "*have you experienced shortness of breath today*?" (five-point Likert scale: 1 = none, 5 = extreme) and "*have you experienced more shortness of breath than usual*?" (yes/no). To encourage short-term reflection on the first day, participants could freely annotate external variables (e.g. weather, mood, physical activity, etc.) which may have influenced their dyspnea. To promote reflection on the second day, participants could only select predefined external variables using check boxes in place of annotations. Additionally, a question about their dyspnea using the

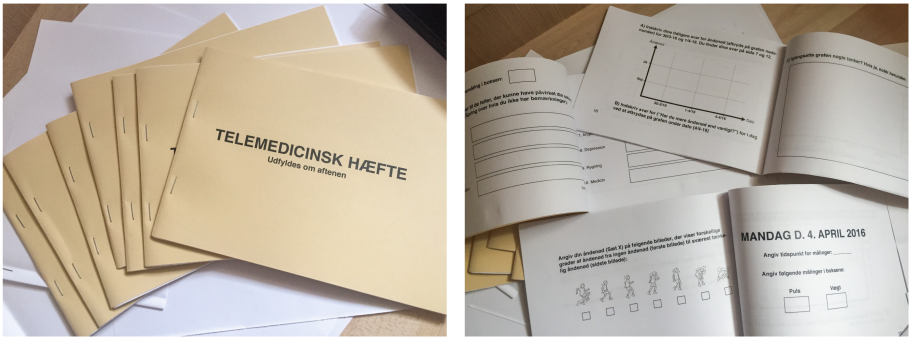
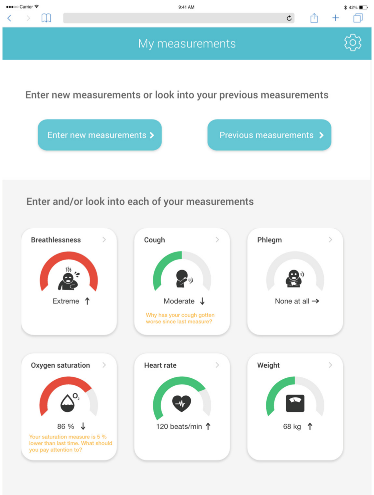


Fig. 1. Workbooks with assignments



Enter new measurements or look into your previous measurements

Enter and/or look into each of your measurements

Breathlessness: Extreme ↑

Cough: Moderate ↓

Why has your cough gotten worse since last measure?

Phlegm: None at all →

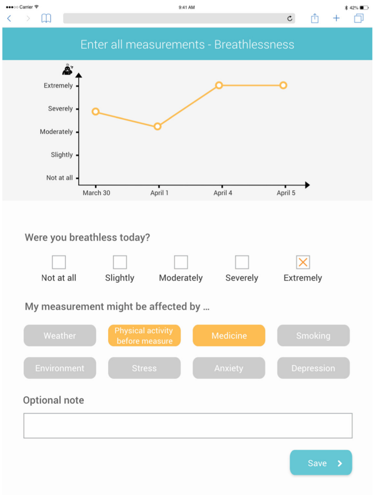
Oxygen saturation: 86 % ↓

Your saturation measure is 5 % lower than last time. What should you pay attention to?

Heart rate: 120 beats/min ↑

Weight: 68 kg ↑

Fig. 2. Overview screen: top - buttons for data entry, bottom - overview of the six measurements' gauges



Were you breathless today?

☐ Not at all ☐ Slightly ☐ Moderately ☐ Severely ☒ Extremely

My measurement might be affected by …

Optional note

Fig. 3. Data entry screen: top - time-series line graph, middle - data entry, bottom - context-relevant variables

seven-point Dalhousie Pictorial Scale was included [33]. To promote long-term reflection on the final day, participants could insert the measurements they collected throughout the use of the diary for blood oxygen levels and the two aforementioned questions into time series graphs. The graphs displayed the participants' recorded measurements on the y-axis and the day along the x-axis. The graphs included a line depicting the recommended health level to trigger reflection in the case of discrepancies between current and target measures. The dashboard prototype consisted of three screens: 1. an overview screen (see Fig. 2), 2. a data entry screen for data collection (see Fig. 3), and 3. a previous measurements screen (see Figs. 4, 5, 6, and 7).

The data entry screen mimicked the diary page for entering pulse, weight, and blood oxygen level but only contained one question: "*were you breathless today?*" (five-point Likert scale: none, to extreme). To help patients' remember previous values during data entry, the data entry screen included a time-series line graph. For each data entry users could add context variables describing factors when the measurement was taken (e.g. being stressed, weather, etc.) to relate them to their measurements and aid in pattern identification.

Completing data entry returned users to the overview page, which encouraged short-term reflections and further data exploration by including reflective questions (e.g. "*Why are you more out of breath than last time you measured?*") alongside a gauge. The six colour coded gauges seen in Fig. 2 indicated the latest measure (arc length) and whether patients were at or below (yellow, red) recommended healthy levels or not (green) to trigger reflections. Trend arrows indicated changes from the penultimate measure (up: improvement, down: worsening).

The comparison page aimed to support long-term reflection and included four different visualisations. The first comparison screen utilized a combined time series graphs

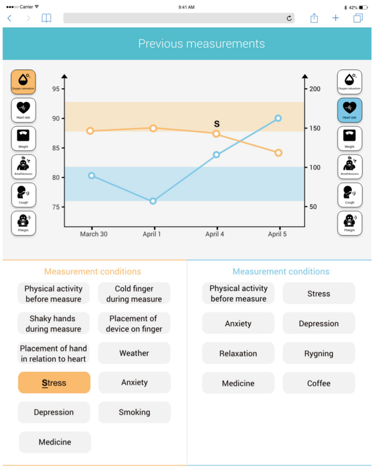


Fig. 4. Comparison screen 1: Top - select measurements to include in the dual axis time series graphs. Bottom - selection of context-relevant variables

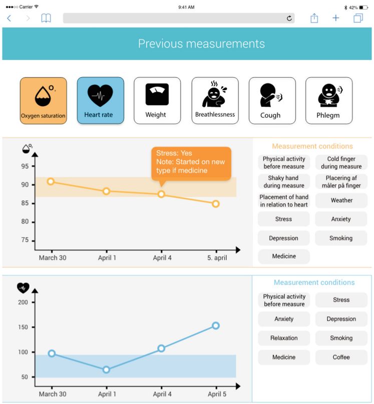


Fig. 5. Comparison screen 2: Top - select measurement. Bottom - time series graphs vertically arranged with context-relevant variables to the right

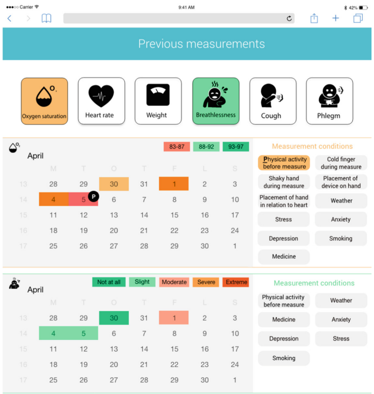


Fig. 6. Comparison screen 3: Top - select measurements. Bottom - Calendar heatmaps with context-relevant variables to the right

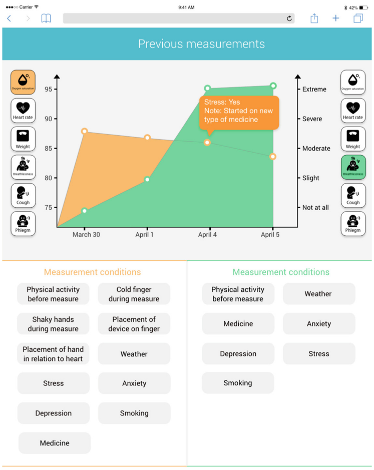


Fig. 7. Comparison screen 4: Top - select measurements to include in the dual axis area graphs. Bottom - select context-relevant variables

to show multiple measurements (see Fig. 4) while the second comparison screen separated and vertically stacked the time series graphs (see Fig. 5). The third comparison screen contained a calendar heatmap to visualise periodic patterns using color shades to indicate daily deviations from recommended levels (see Fig. 6). The fourth comparison screen contained area graphs visualising multiple measurements (see Fig. 7). The visualisations aimed to promote reflections on patients' health trends and increase awareness of possible worsening conditions. The visualisations allowed for comparisons of multiple measures and context-relevant variables to trigger reflection on how measures impacted each other. Recommended levels on visualisations aimed to increase awareness of discrepancies and trigger reflection.

4.1 Participants and Method

We recruited participants through a local hospital. A nurse involved in the currently used COPD telehealth service made initial contact over the phone and upon agreeing provided us with their contact information. We introduced them to the details of the study over the phone, and after consenting, the participants received an information letter and consent form prior to an initial in-person meeting.

Five COPD patients, two men (age M: 64.5) and three women (age M: 66.8), participated in the study. Three used supplemental oxygen (P3, P4, P5) and all lived in their own homes with spouses, except for P3 who lived alone. P5 had a speech disorder so her spouse (P5S) spoke on their behalf. All patients had multiple co-morbidities (asthma, diabetes, heart disease, etc.). We required that all participants were currently using a mandated self-tracking telehealth application - in this case ambuflex - to collect data for their healthcare professionals. Ambuflex did not provide the patients with any feedback on their condition.

The participants received the paper diary one week prior to the session evaluating the dashboard. On the day of its evaluation, we collected the data from the patient diaries and updated the prototype using the patients' measurements to allow reflections on their own data. First, patients provided feedback on the diary in an unstructured interview. Afterwards, the patients went through each dashboard screen while completing tasks such as entering blood oxygen saturation data, comparing measures, etc. in a think-aloud manner [43], followed by a short debrief interview. The entire evaluation was limited to less than an hour and required minimal physical activity from the participants as co-designing with COPD patients using generative techniques (e.g. post-it notes and sketching activities) was often too demanding for them [10]. Nadarajah et al. similarly experienced that COPD patients within an hour of interviewing experienced breathing difficulties requiring a slow pace and long breaks [29].

4.2 Results

Collecting Data. Patients varied in terms of when and how they took measures. Three patients had specific schedules for recording data (e.g. always before breakfast) while others recorded inconsistently. We identified two types of patients in this study: Passive patients (P1 and P3) who took the role of data providers without further reflections and proactive patients (P2, P4, and P5) who engaged and reflected on their data. Passive patients lacked knowledge on how the context of recording data influenced their measures. For example, they did not understand how a cold finger when measuring oxygen saturation reduced the validity of their data. Alternatively, proactive patients ensured taking measures under comparable conditions and noted additional variables relevant for their measures (e.g. mood, supplemental oxygen, etc.). P2 and P5S asked for guidelines on taking reliable and valid measures: *"For some measures, an explanation would be good. For example, do not take measure if this and that"* (P5S).

All patients preferred higher granularity options (e.g. Likert scales) when rating symptoms as opposed to binary scales (yes/no): *"How much is a no? If we say yes or no to the hospital, they still do not know what we are thinking. They'll call us and we'll have to explain the severity"* (P5S). However, rating symptoms on a Likert scale without a baseline (*"Did you feel breathless today?"*) caused difficulties for four patients as the severity of their symptoms varied during the day: *"I have been through all of the provided options that day. How do you want me to answer that?"* (P4) and *"If she [P5] is not more breathless than yesterday, then we'll just submit a no [when as asked if they feel breathless today]"* (P5S). They also had different perceptions of how to rate the severity of their symptoms: *"I base that [rating] on when I'm at my best"* (P3) and *"usual is when it is an ordinary day"* (P4).

Reflecting on the Dashboard. While patients diligently recorded data, only P5 responded to the reflective exercises in the paper diary. When given the dashboard, the passive patients felt unmotivated to reflect on the visualisations, reflective questions, or data history: *"I do not care what my status is. I just submit the data. I do not walk around and think every day about how I am feeling"* (P3). In contrast, proactive patients reflected on their data to determine actions to take: *"I'm more concrete. Where am I right now and what can I do about it?"* (P4). However, four patients found the data too complicated to reflect on: *"I do not know what they use it for, the scales they use and the language. I do not understand it. I count on them [nurses, doctors, etc.] to react if there is anything"* (P2). These participants relied on healthcare workers to review the self-tracked data and explain any negative results: *"I have a nurse who is good at keeping an eye on me"* (P4). Additionally, passive participants did not feel the need to reflect on their data due to relying on their healthcare workers: *"I do not need it [access to history data]. If it [measure] is too low, they call and ask me why"* (P1).

The overview screen helped patients understand their health status: *"[I can] quickly see if it [a measure] is going up or down"*. Additionally, according to three patients, the gauges of recommended healthy levels for each measure simplified the process of identifying unhealthy measures and taking actions to improve: *"If it starts to go over here [below recommended], we have to do something"* (P5S) and *"[my health] is not that bad if I keep it [a measure] above that lower threshold"* (P2). However, three patients felt demotivated by comparing their data to thresholds when they understood the consequences of falling outside healthy ranges but not how to avoid this: *"I prefer not to be told in the morning that I'm gonna get an awful day"* (P4) and *"It's OK if it's just a single measure [outside of health thresholds], but if it is constant, I would start thinking it [my health] is going away fast now"* (P2). To avoid this, they suggested personalizing the dashboard e.g. setting recommended levels based on the severity of their condition.

Active patients valued looking at line graphs showing their health over time in the overview screen: *"This gives more information about me (...) it's nice to be able to go* *back. Is it better than 14 d ago?"* (P2). However, one proactive patient needed a purpose to reflect to gain any benefits from history data: *"There might be days where I sit with it and have an idea about what I'm looking for, which might trigger some thoughts"* (P4). Passive patients did not see the benefits from reflecting on the data themselves: *"this [overview screen] is only for people who have to sit and analyse the numbers"* (P1). The reflective questions in the overview screen did not trigger any reflections in the passive patients who ignored the questions. While proactive patients did answer the reflective questions, they did not find an answer using the data but instead proposed one from their previous knowledge (e.g. to the question *"Why are you coughing more than last time you measured?"*, P4 answered: *"Right now it is likely because I talk too much"*).

When looking at the comparison screens, four patients preferred the dual axis time-series line (see Fig. 4), which simplified finding discrepancies and relations between measures: *"you can have them [measures] together and see how they affect one another"* (P2). Three patients felt that the calendar heat-map in Fig. 6 provided a simple explanation of their health over time: *"it [Fig. 6] is the one I understand the quickest"* (P4).

In summary, patients struggled to understand how to collect measures under reliable circumstances and interpreted questions regarding the severity of their symptoms differently. Patients' initial thoughts on the dashboard varied depending on whether they had a passive or proactive attitude towards their treatment. While passive patients felt the visualisations offered little benefits for them, proactive patients liked to reflect using the time-series line charts and calendar heat-maps. However, Study 1 only provided insight into patients' initial and potentially - to please the researchers - biased thoughts. A two-week follow-up study investigated attitudes about and behaviour changes from using a self-tracking dashboard over time.

5 Study 2 - Evaluating Reflection During Use

Based on the results from Study 1, we redesigned the dashboard to only include the time-series line chart. In this study, we explored how using the new dashboard over a 14-day period impacted reflection among COPD patients and their activities in managing their condition.

5.1 Prototype Redesign

We implemented the dashboard as a web-application accessed through a tablet or phone. An introductory dialogue box provided information on how to measure data under comparable conditions and report context-related variables that can impact measurements. Reflective questions targeted patients' overall health instead of specific measures (e.g. *"Have you previously been able to improve your measures? How?"*). Some reflective questions aimed to increase patients' awareness of symptom changes (e.g. *"You have multiple measures showing red/yellow. Have you explored what your measures might have been affected by?"*). A setting allowed patients to adjust thresholds indicating recommended levels for each measure to their own preference. The visualisations on the comparisons screen were limited to the dual y-axis time-series line graph seen in Fig. 4 as patients preferred it most in Study 1.

5.2 Participants and Method

Five COPD patients, two male (age M: 71.5) and three female (age M: 75), participated in the study - none of whom participated in Study 1. Patients were diagnosed with COPD between seven and 25 years ago (M: 12) and experienced either moderate, severe, or very severe COPD. Two patients (P3, P5) used supplemental oxygen and three (P3, P4, P5) suffered from multiple co-morbidities (diabetes, osteoporosis, and fibromyalgia). P4 reported colour blindness but could distinguish between the colour used in the dashboard. Using the same procedure as Study 1, we recruited participants through a local hospital with the help of nurses.

During an initial meeting, the patients received both a written and verbal explanation of the study once more and consented that no healthcare professional would review the data collected through the study's equipment. They received a self-tracking kit consisting of a pulse oximeter, weight scale, diary containing a template to track measurements, and a tablet with internet access. Patients measured and recorded oxygen saturation, pulse, weight, self-reported dyspnea, cough, and phlegm into the dashboard prototype for 14 d. We encouraged the patients to also record their measures in the diary but made it optional to reduce the effort required from the patients. We suggested recording measures three times a week and asked those currently using self-tracking to use our dashboard on days on which they did not use their existing system.

A facilitator instructed the patients on the use of the system such as: Opening the application, submitting data, accessing previous measures, and adjusting settings. For each session in which the dashboard was used the system automatically logged all user interactions: 1) time spent on each screen and total time per session, 2) where, what, and how many times the screen was clicked. The data was anonymised and stored on a secured server.

After 14 d, we conducted semi-structured interviews in participants' homes. Each interview lasted between 53 min and 1 h and 45 min revolving around: COPD-related activities for managing disease, context of use, and comparisons with previous self-tracking methods. We prepared screenshots of patients' dashboards showing events of interest (e.g. worsening or improvement in measures between two days) and scanned the patients' diaries for significant events before the interview. The resulting interview data were analysed using grounded theory methods [14].

5.3 Results

The patients entered data 4–5 times during the 14 d except one patient who did so nine times. Usage sessions took on average nine and a half minutes. The longest came from P4 who used 32 min to enter measures, answer reflective questions, and interact with visualisations. The shortest (3 min), consisted of only entering measures. Most sessions consisted of patients spending approximately 75% of their time on the data entry screen, afterwards they used the overview and comparison screens. While three patients only viewed these screens after entering data, two patients consulted the dashboard without data entry. Four patients acted on reflective questions on the overview screen by exploring their measures to identify factors that could explain their negative health.

Five themes emerged from the analysis: 1) Motivation for reflection and system use, 2) using measures as health status indicators, 3) feeling empowered in everyday life, 4) gaining self-knowledge, and 5) becoming motivated to self-improve.

Motivation for Reflection and System Use. Four patients cited their agreement to participate in this study as motivation for using the dashboard while the last felt motivated by reflecting on their health and taking actions to improve it. Similar to Study 1, we classified three patients as proactive (P1, P3 and P4) and two as passive (P2 and P5).

Similar to Study 1, passive patients lacked the knowledge to improve their health through reflection: *"we can not do anything except measure"* (P2) and *"if the bright minds can not make sure that I get better, then neither can I do anything about it"* (P5). These patients doubted that they could improve their health long-term: *"I do not worry about things that I can not change"* (P2). However, they still reflected on the dashboard for short-term improvements such as adjusting their supplemental oxygen levels.

Using Measures as Health Status Indicators. All patients reflected on their health on days of bad health: *"[when] I actually feel good, I do not worry about how I felt yesterday"* (P3). Four patients did not reflect on past data when they felt well to avoid remembering bad days: *"that's not something I walk around and think about. Life gets too strenuous if you walk around and think about that [bad days in past]"* (P1). However, four patients reflected on their data to explore possible reasons for why they felt unwell: *"if I do not feel like everything is fine, I might start thinking why (...) it depends on how I am feeling"* (P4).

All patients reported that self-tracking and reflecting increased their awareness of how they felt: *"I start noticing three times a week, how am I feeling right now?"* (P4) and used their pulse oximeter to check their current status and took action to improve their condition (e.g. performed breathing exercises after measuring low oxygen saturation). According to them the dashboard provided a quick and simple overview of their health for short-term reflections: *"it [the dashboard] is a measure of one's symptoms (...) altogether it of course becomes how you are feeling"* (P1). Proactive patients used the dashboard to identify reasons for feeling unwell *"you can not always go to the doctor and learn about your status and why you feel that way (...) you can do that here [dashboard]"* (P3).

Questioning and Gaining Self-Knowledge. Proactive patients gained insights by asking themselves questions and increasing their awareness of what caused their symptoms to intensify. For example, the reflective questions in combination with annotating measures with context variables triggered reflection in proactive patients: *"with dyspnea, I had not thought there could be other [reasons]. I just had breathlessness, done. (...) suddenly I realized how much I was affected by the heat (...) it happened when I sat with the system and those questions asking 'why?'"*. Annotating measures with context variables supported evaluating different causal explanations: *"I have started thinking about it (...) I think, 'no it's not that [stress]', 'Talk? No I haven't talked today' and then I think 'it's the weather'"* (P3). Some proactive patients would like the dashboards to identify and highlight the important contexts which influenced their symptoms. These patients reflected on previous days to identify changes that effected their health: *"I become very conscious about, how did I feel yesterday? Do I also feel like that today? What caused that?"* (P4).

Empowered Through Reflection. Through reflections, three proactive patients learned how previously ignored measurements impacted their condition in everyday life. These patients felt empowered by self-tracking and gained agency over their health: *"I thought that is just how it is. You give up a little and get tired of it [COPD] (...) without doing anything about it, nobody says anything, but this [the system] does. It makes you aware of the situation (...) My doctor always told me that it [their negative mentality] is all because of my condition. The system makes me think that he is not right"* (P3). Two patients felt empowered by identifying correlations between their symptoms and contextual variables (e.g. warm weather may result in breathing difficulties) which they used to inform their actions: *"now I can make up my mind beforehand [whether to go outside in the heat], because I know how it will end"* (P3). Similarly, these patients aimed to keep their measures within recommended levels and felt safer knowing that their health has not deteriorated to dangerous levels: *"I'm on the right track then"* (P3). Three patients felt empowered by preventing family members from witnessing bad symptoms: *"I can become unsure about how I am feeling.. (...) I do not want to expose my husband and daughter unnecessarily [frightening events] (...) I learn more about that now, so that I do not expose them"* (P3) and *"I have to be self-centred (...) I have to do things right for myself and in time, so that I also treat others right"* (P4).

Becoming Motivated to Self-Improve. Active patients used the dashboard to set goals for improving their health which motivated them to seek new knowledge that can aid their goal: *"I've tried to acquaint myself with BMI because I wanted to have a goal to follow. [because] I wondered about the arrows [in the system]"* (P4). One proactive patient learnt about the severity of their weight problem and gained awareness about the need to improve: *"I have not thought about it before, but when you suddenly get it in writing (...) being confronted with it, I have to do something about it (...) it's for my own good"* (P3). That patient used the dashboard to track their progress as they tried to improve their diet: *"that's about getting better at using the device [tablet with the dashboard]. Not just saying, 'oh, you are running into a pneumonia, now you have to use it, it's about using it [the dashboard] several times a day"* (P3).

Three patients stated that the overview screen containing the colour indicators and arrows provided them with a concrete goal to pursue: *"I want all of them [days] to be green and that things are making progress"* (P3) and *"when the arrows are pointing down I assume it is not so good, that's the wrong way"* (P4). For example, one proactive patient used the dashboard to help reduce medication intake, which had been a struggle despite the doctor's encouragement: *"They [doctors] had difficulties easing me off because I have had high doses for so many years (...) but this time I thought now you have to stop (...) I did, I needed some days and then it was over"* (P3). Two patients stated that the dashboard should contain advice on actions to take that can improve their health. Specifically, they wanted advice tailored to their own goals and health problems: *"to get help when you also have diabetes, that would be nice"* (P3).

6 Discussion

The proactive patients in our study felt motivated to self-track their symptoms, unlike the chronically ill patients in larger scale studies, who, however, could not review or interact with their entered data [1,41]. Our prototype supported such activities and as previously hypothesized [1,41], the ensuing reflections empowered the proactive patients to reduce their symptoms over time boosting motivation for tracking. Using reflective "why" questions in our dashboard overviews to highlight measurements that had changed since last time, prompted comparing the contextual annotations to the change in measurements illustrated in time series graphs. This triangulation of reflected questions, context annotations, and time series graphs proved instrumental in providing the proactive patients with agency, similar to findings with healthy self-trackers [13,20]. However, the passive patients lacked motivation to self-track as they felt unable to reflect and doubted the application's efficacy to empower them to that end. Our proactive patients shared this doubt until they could identify concrete actions and improvements (see [19]).

Contrasting previous studies [1,25,32,40,41], all our patients tracked their symptoms with some proactive patients submitting more data than asked of them. They attributed much of their motivation to start using the application to the social contract they had entered by participating in the study. However, the proactive patients' increased sense of agency additionally motivated them to self-track, supporting suggestions from previous studies [1,6,20]. The patients had to identify worsened health and reflect on visualisations - a common approach in this application type (see [25,26]). Similar to stroke patients, all our patients had a bias against reflecting on data reminding them of bad health and either disregarded the results [15] or temporarily suspended self-tracking [1]. Once our proactive patients learned to reflect and identify actions to potentially improve their health, they analysed their negative health data in relation to their contextual annotations (see [1,17]). The proactive patients used their newfound insight to improve their condition, e.g. by not going outside during bad weather, adjusting oxygen supplements, and reducing their condition's impact on their social lives - similar to COPD patients in previous studies [17,32].

However, this gained agency did not come without risks. Patients could reach incorrect conclusions through their biases, assumptions, and misunderstandings, which is a common concern in interfaces relying on users to create insights from data visualisations [37]. For example, our patients preferred dual y-axis time series charts to compare different measure(s) to identify potential reasons for changes in their health. However, these charts are known to suggest correlations where none might exist [4]. This points to more general design dilemmas pointed out by Correll [8] regarding the degree of agency users should hold to be empowered while protecting them from arriving at potentially spurious conclusions. This was further exemplified by the articulated need for automated analysis from proactive patients, who felt burdened by searching through every variable presented in the dashboard. Adaptive dashboard could automatically highlight pertinent variables relevant to goal setting and problems. How much guidance should these systems provide and should they limit users' explorations? Other future work should investigate how novel telehealth interfaces (e.g. virtual assistants) can establish social contracts with patients as well as help with on-boarding, measuring health reliably, and reflecting on results using visualisations, reflective questions, and contextual annotations.

While our findings are based on a smaller number of patients than studies relying on short interaction sessions [15,22,24], our numbers are similar to studies investigating usage of telehealth dashboards for longer, e.g. fortnight, periods [25,26]. Our participants had all received their diagnoses multiple years ago, adapted their lives to accommodate their condition, and self-tracked prior to the study. Therefore, our results may not apply to novice users, who still adapt to their condition and lack self-tracking experience.

7 Conclusion

When used longer term, tablet-based telehealth dashboards utilizing reflective "why" questions to highlight change in measurements, and contextual annotated time-series graphs can encourage proactive patients to reflect, self-track, and improve their quality of life through an increased understanding of their health. However, patients needed knowledge about measuring health parameters and how to follow up results indicating poor or declining health with concrete actions to reflect. visualisations of tracked data and reflective questions might not motivate patients, who understand their role to be mere data providers. Future studies should investigate how to create adaptive dashboards that can promote reflections and actions relevant to each patients' circumstances and how to navigate the inherent design dilemmas between the empowerment of vulnerable users and protecting from taking action based on incorrectly drawn conclusions in interventions that promote taking an active role.

Acknowledgements. We would like to thank Britta Ravn from the Center for Telemedicin and Silkeborg Regional Hospital for helping with participant recruitment and the anonymous reviewers for their constructive feedback to improve the final version of the paper. This article/publication is based upon work from the COST Action IRENE (CA18118), supported by COST (European Cooperation in Science and Technology).

References

1. Ancker, J.S., Witteman, H.O., Hafeez, B., Provencher, T., Van de Graaf, M., Wei, E.: You get reminded you're a sick person: Personal data tracking and patients with multiple chronic conditions. J. Med. Internet Res. **17**(8), e202 (2015). https://doi.org/10.2196/jmir.4209
2. Atkins, S., Murphy, K.: Reflection: a review of the literature. J. Adv. Nurs. **18**(8), 1188–1192 (1993)
3. Brandt, C.L.: Study of older adults' use of self-regulation for COPD self-management informs an evidence-based patient teaching plan. Rehabil. Nurs. **38**(1), 11–23 (2013). https://doi.org/10.1002/rnj.56

4. Brath, R., Hagerman, C., Sorenson, E.: Why two y-axes (y2y): a case study for visual correlation with dual axes. In: 2020 24th International Conference Information Visualisation (IV), pp. 38–44 (2020). https://doi.org/10.1109/IV51561.2020.00016
5. Chen, Y., Pu, P.: HealthyTogether: exploring social incentives for mobile fitness applications. In: Proceedings of the Second International Symposium of Chinese chi, pp. 25–34 (2014). https://doi.org/10.1145/2592235.2592240
6. Choe, E.K., Lee, N.B., Lee, B., Pratt, W., Kientz, J.A.: Understanding quantified-selfers' practices in collecting and exploring personal data. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, pp. 1143–1152. ACM (2014). https://doi.org/10.1145/2556288.2557372
7. Chung, C.F., Cook, J., Bales, E., Zia, J., Munson, S.A.: More than telemonitoring: Health provider use and nonuse of life-log data in irritable bowel syndrome and weight management. J. Med. Internet Res. **17**(8), e203 (2015). https://doi.org/10.2196/jmir.4364
8. Correll, M.: Ethical dimensions of visualization research. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM (2019). https://doi.org/10.1145/3290605.3300418
9. Cuttone, A., Petersen, M.K., Larsen, J.E.: Four data visualization heuristics to facilitate reflection in personal informatics. In: Stephanidis, C., Antona, M. (eds.) UAHCI 2014. LNCS, vol. 8516, pp. 541–552. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07509-9_51
10. Das, A., Bøthun, S., Reitan, J., Dahl, Y.: The use of generative techniques in co-design of mHealth technology and healthcare services for COPD patients. In: Marcus, A. (ed.) DUXU 2015. LNCS, vol. 9188, pp. 587–595. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20889-3_54
11. Epstein, D.A., Ping, A., Fogarty, J., Munson, S.A.: A lived informatics model of personal informatics. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 731–742. ACM (2015). https://doi.org/10.1145/2750858.2804250
12. Feustel, C., Aggarwal, S., Lee, B., Wilcox, L.: People like me: designing for reflection on aggregate cohort data in personal informatics systems. Proc. ACM Interact. Mobile, Wearable Ubiquit. Technol. **2**(3), 1–21 (2018). https://doi.org/10.1145/3264917
13. Fleck, R., Fitzpatrick, G.: Reflecting on reflection: framing a design landscape. In: Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction, pp. 216–223. ACM (2010). https://doi.org/10.1145/1952222.1952269
14. Glaser, B.G., Strauss, A.L.: The Discovery of Grounded Theory: Strategies for Qualitative Research. Routledge, UK (1967)
15. Hougaard, B.I., Knoche, H.: Telling the story right: how therapists aid stroke patients interpret personal visualized game performance data. In: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, pp. 435–443 (2019). https://doi.org/10.1145/3329189.3329239
16. Isaacs, E., Konrad, A., Walendowski, A., Lennig, T., Hollis, V., Whittaker, S.: Echoes from the past: how technology mediated reflection improves well-being. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1071–1080. ACM (2013). https://doi.org/10.1145/2470654.2466137
17. Kaptain, R.J., Helle, T., Kottorp, A., Patomella, A.H.: Juggling the management of everyday life activities in persons living with chronic obstructive pulmonary disease. Disabil. Rehabil. **44**(14), 3410–3421 (2021). https://doi.org/10.1080/09638288.2020.1862314
18. Khare, S.R., Vedel, I.: Recall bias and reduction measures: an example in primary health care service utilization. Family Pract. **36**(5), 672–676 (2019). https://doi.org/10.1093/fampra/cmz042

19. Koff, P., Jones, R.H., Cashman, J.M., Voelkel, N.F., Vandivier, R.: Proactive integrated care improves quality of life in patients with COPD. Eur. Respir. J. **33**(5), 1031–1038 (2009). https://doi.org/10.1183/09031936.00063108
20. Li, I., Dey, A., Forlizzi, J.: A stage-based model of personal informatics systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 557–566. ACM (2010). https://doi.org/10.1145/1753326.1753409
21. Li, I., Dey, A.K., Forlizzi, J.: Understanding my data, myself: supporting self-reflection with ubicomp technologies. In: Proceedings of the 13th International Conference on Ubiquitous Computing, pp. 405–414. ACM (2011). https://doi.org/10.1145/2030112.2030166
22. Liu, L.H., et al.: Patient and clinician perspectives on a patient-facing dashboard that visualizes patient reported outcomes in rheumatoid arthritis. Health Expect. **23**(4), 846–859 (2020). https://doi.org/10.1111/hex.13057
23. MacLeod, H., Tang, A., Carpendale, S.: Personal informatics in chronic illness management. In: Proceedings of Graphics Interface 2013, pp. 149–156. Canadian Information Processing Society (2013)
24. Martinez, W., Threatt, A.L., Rosenbloom, S.T., Wallston, K.A., Hickson, G.B., Elasy, T.A.: A patient-facing diabetes dashboard embedded in a patient web portal: design sprint and usability testing. JMIR Hum. Factors **5**(3), e26 (2018). https://doi.org/10.2196/humanfactors.9569
25. Meyer, J., Beck, E., Wasmann, M., Boll, S.: Making sense in the long run: long-term health monitoring in real lives. In: 2017 IEEE International Conference on Healthcare Informatics (ICHI). IEEE (2017). https://doi.org/10.1109/ichi.2017.11
26. Meyer, J., Kay, J., Epstein, D.A., Eslambolchilar, P., Tang, L.M.: A life of data: characteristics and challenges of very long term self-tracking for health and wellness. ACM Trans. Comput. Healthc. **1**(2), 1–4 (2020). https://doi.org/10.1145/3373719
27. Mols, I., van den Hoven, E., Eggen, B.: Technologies for everyday life reflection: illustrating a design space. In: Proceedings of the TEI 2016: Tenth International Conference on Tangible, Embedded, and Embodied Interaction, pp. 53–61. ACM (2016). https://doi.org/10.1145/2839462.2839466
28. Muller, L., Divitini, M., Mora, S., Rivera-Pelayo, V., Stork, W.: Context becomes content: sensor data for computer-supported reflective learning. IEEE Trans. Learn. Technol. **8**(1), 111–123 (2015). https://doi.org/10.1109/TLT.2014.2377732
29. Nadarajah, S.G., Pedersen, P.W., Hougaard, B.I., Knoche, H.: Am i coughing more than usual? Patient reflections and user needs on tracking COPD data in a telehealth system. In: Proceedings of the 4th International Workshop on Multimedia for Personal Health & Health Care, pp. 2–8. HealthMedia 2019, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3347444.3356237
30. Oh, J., Lee, U.: Exploring UX issues in quantified self technologies. In: 2015 Eighth International Conference on Mobile Computing and Ubiquitous Networking (ICMU), pp. 53–59. IEEE (2015)
31. Patel, R.A., Klasnja, P., Hartzler, A., Unruh, K.T., Pratt, W.: Probing the benefits of real-time tracking during cancer care. In: AMIA Annual Symposium Proceedings, vol. 2012, p. 1340. American Medical Informatics Association (2012)
32. Pedone, C., Lelli, D.: Systematic review of telemonitoring in COPD: an update. Pneumonol. Alergol. Pol. **83**(6), 476–484 (2015). https://doi.org/10.5603/PiAP.2015.0077
33. Pianosi, P.T., Huebner, M., Zhang, Z., McGrath, P.J.: Dalhousie dyspnea and perceived exertion scales: Psychophysical properties in children and adolescents. Respir. Physiol. Neurobiol. **199**, 34–40 (2014). https://doi.org/10.1016/j.resp.2014.04.003
34. Puussaar, A., Clear, A.K., Wright, P.: Enhancing personal informatics through social sensemaking. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 6936–6942 (2017). https://doi.org/10.1145/3025453.3025804

35. Pelayo, V.R.: Design and Application of Quantified Self Approaches for Reflective Learning in the Workplace. KIT Scientific Publishing, Germany (2015)
36. Rogers, R.R.: Reflection in higher education: a concept analysis. Innov. High. Educ. **26**(1), 37–57 (2001). https://doi.org/10.1023/A:1010986404527
37. Sacha, D., Senaratne, H., Kwon, B.C., Ellis, G., Keim, D.A.: The role of uncertainty, awareness, and trust in visual analytics. IEEE Trans. Vis. Comput. Graph. **22**(1), 240–249 (2015). https://doi.org/10.1109/TVCG.2015.2467591
38. Sorenson, E., Brath, R.: Financial visualization case study: correlating financial timeseries and discrete events to support investment decisions. In: 2013 17th International Conference on Information Visualisation, pp. 232–238. IEEE (2013). https://doi.org/10.1109/IV.2013.31
39. Sparring, V., Granström, E., Sachs, M.A., Brommels, M., Nyström, M.E.: One size fits none-a qualitative study investigating nine national quality registries' conditions for use in quality improvement, research and interaction with patients. BMC Health Serv. Res. **18**(1), 802 (2018). https://doi.org/10.1186/s12913-018-3621-9
40. Ure, J., et al.: Piloting tele-monitoring in COPD: a mixed methods exploration of issues in design and implementation. Prim. Care Respir. J. **21**(1), 57–64 (2011). https://doi.org/10.4104/pcrj.2011.00065
41. Verdezoto, N., Grönvall, E.: On preventive blood pressure self-monitoring at home. Cogn. Technol. Work **18**(2), 267–285 (2015). https://doi.org/10.1007/s10111-015-0358-7
42. Vázquez-Ingelmo, A., Garcia-Peñalvo, F.J., Therón, R.: Information dashboards and tailoring capabilities - a systematic literature review. IEEE Access **7**, 109673–109688 (2019). https://doi.org/10.1109/ACCESS.2019.2933472
43. Wright, P.C., Monk, A.F.: The use of think-aloud evaluation methods in design. SIGCHI Bull. **23**(1), 55–57 (1991). https://doi.org/10.1145/122672.122685
44. Zuckerman, O., Gal-Oz, A.: Deconstructing gamification: evaluating the effectiveness of continuous measurement, virtual rewards, and social comparison for promoting physical activity. Pers. Ubiquit. Comput. **18**(7), 1705–1719 (2014). https://doi.org/10.1007/s00779-014-0783-2

Design of a Social Chatbot with Gamification for user Profiling and Smoking Trigger Detection

Kyana Bosschaerts[1](✉), Jeroen Stragier[2], Bram Steenwinckel[1],
Lieven De Marez[2], Sofie Van Hoecke[1], and Femke Ongenae[1]

[1] IDLab, Ghent University-imec, Technologiepark 126, 9052 Gent, Belgium
`kyana.bosschaerts@ugent.be`
[2] MICT, Ghent University-imec, Miriam Makebaplein 1, 9000 Gent, Belgium
`jeroen.stragier@ugent.be`

Abstract. User profiling is essential to help smokers quit smoking, but filling out a questionnaire is tedious and therefore, a lot of smokers drop out even though this personalisation can help them greatly in their journey to quit smoking. In this project a chatbot is designed that acquires the necessary data for personalisation through games. Combined with smoking event registrations, the application can detect smoking triggers. 12 participants tested the chatbot for 15 d by talking to it and registering their smoking events. Results show that the chatbot reaches the requirements of a social chatbot, gathers for most games good quality data, and detected smoking triggers are accurate, making the chatbot a great alternative for smokers with an interest in games.

Keywords: Smoking cessation · Social chatbot · Gamification · User profiling · Behavioral change interventions · Digital health

1 Introduction

According to the WHO, modifiable behaviors, such as tobacco use and physical inactivity, cause 80% of non-communicable diseases (NCDs), including cardiovascular disease, type 2 diabetes and cancer [21]. NCDs kill 41 million people each year, equivalent to 71% of all deaths globally. 70–85% of the medical budget of OECD countries is used to treat NCDs. Preventive medicine, focusing on measures to modify a patient's behavior in order to prevent diseases, has the potential to reduce NCD prevalence, improve quality of life and to reduce healthcare costs. Within this research, we specifically focus on encouraging smoking cessation. Despite smoking cessation program development and policy measures in the past decades, still almost 1 in 5 Belgians is a smoker [18]. Tobacco use accounts for over 14.000 premature deaths in Belgium every year. The direct cost to healthcare in Belgium is estimated at 615 million euro and indirect costs, such as absenteeism, are another 746 million. These numbers illustrate the societal and economic importance of designing engaging smoking cessation programs.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 636–650, 2023.
https://doi.org/10.1007/978-3-031-34586-9_42

Digital Health Behavior Change Interventions (DBCI) are being developed that complement the face-2-face coaching using mobile applications and wearables [22]. They aim to empower individuals by collecting and visualizing vast amounts of behavioral data in a comprehensive manner and incorporating behavior change techniques to promote a healthier lifestyle, e.g. goal-setting, social support, gamification (rankings and rewards). It has been demonstrated that DBCIs have higher effectiveness when data insights, social, challenges and motivational messages are tailored to the profile of the user [13]. Personalisation is thus an import prerequisite to achieve effective smoking cessation DBCIs [9]. An important part of enabling smoking cessation, is accurately identifying the smoking triggers of a person, i.e. specific contexts, times, locations, social behavior or mindsets that trigger the person to smoke. This allows the DBCI to intervene at the most critical moments for this particular smoker, so that appropriate measures can be taken to prevent a relapse [8].

Today, profile and trigger information is mainly gathered through questionnaires, or by using a smartphone or a wearable, e.g. for collection of location data [19,20]. The problem with gathering personal information through questionnaires is that users oftentimes tend to lose interest after only a few questions [12]. Consequently, a significant proportion of users drop out of the questionnaire. Even though the personalisation would add a lot of value to the user in the end, the process of filling in the information makes the whole endeavour for them not worth it. Sensors that directly detect the information needed for profiling without explicit input from the smoker can partially solve the issue. However, the disadvantages is that these sensors, e.g. wearables, are often expensive, and that they require advanced data analysis to make sense of the data. Moreover, a lot of information is inaccurate because of situational deviations of normal behaviour, a lack of registrations, difficulty to assess how the data should be interpreted within a particular context, or mismeasurements [14].

We therefore propose a social chatbot that gathers the necessary data through games as a solution to the problems regarding the current approaches. Smokers can play a game while in the background the application and the chatbot processes their data, creates their user profile, and detects their smoking triggers. In order to motivate the smoker to play more games, and thus create an opportunity for the chatbot to gather more data, gamification is used [11]. This method of information gathering can replace questionnaires while making it fun for the user to provide their personal information.

This paper is structured as follows. First, the related work is described in Sect. 2. Section 3 contains the design of the chatbot, while Sect. 4 deals with the study set-up. Last of all, the results are given in Sect. 5, while the paper concludes in Sect. 6.

2 Related Work

A few chatbots used in the healthcare already exist. A mental health chatbot reduces symptoms of stress, depression, and anxiety significantly [3]. The higher the engagement with the chatbot is, the more the symptoms of anxiety and depression lower. Another mental health chatbot Tess is designed to act like a therapist that has different modules that correspond to different types of treatment modalities for depression [5]. These different modules bring different levels of engagement according to the length, complexity and style of the question asked. Furthermore, also other chatbots exist that focus more on promoting healthy lifestyles with as advantages that this can reach a broad audience and that automated personalised messages are possible [6].

In the field of smoking cessation, some studies focus on intervention, and while these do not use chatbots, they indicate that the use of one would be beneficial for behavioral change. None of these apply a chatbot yet either [17]. One study uses short message service (SMS) texting to help smokers quit and prevent relapse. Here participants indicate that this texting should be personalised and interacting, thus preferring a chatbot above the one-way method in the study. Moreover, another study indicates that just-in-time interventions may result in similar outcomes as in-person counseling [7]. In this paper smoking triggers are detected, while in another study known smoking triggers are used to personalize an intervention [8]. This results in significantly greater reductions in urges than just general messages.

In the case of chatbots in smoking cessation, a lot of studies focus on intervention for behavioral change while not giving any attention towards the gathering of the great amount of data needed for personalisation. A chatbot already exists that supports the smoker in their journey to quit smoking [10]. This chatbot is used in combination with a popular smoking cessation app - the Smoke Free app. As far as the limited information allows to discover true features of the Smoke Free app used in this study without paying for it, it seems to be limited to a chatbot without engaging games to discover triggers, complementing thus the questionnaires but not being able to replace them. The work however illustrates the need for more engaging interaction with people wanting to stop smoking. Another chatbot elicits reflection in smokers by allowing open answers to questions and using natural language processing to adequately respond to the given answer [1]. Subsequently, it also identifies smoking reasons of smokers.

Gamification in the healthcare often appear as serious games without any involvement of a chatbot. Serious games educate the players about a particular aspect of their health, or try to help a person by giving the player advice through the game. A study compared some of these serious games targeted at changing the behaviour of smokers and concluded that the games positively affected smoking-related outcomes [4].

In summary, increased engagement and personalization in smoking cessation programs is required to allow for effective DBHI, and multiple studies already point towards a chatbot as a possible solution. Even though some interventions help smokers to prevent relapse, still smoking triggers have to be known first before an intervention can be staged. Gathering this profile information is done through questionnaires and sensors, leading user to disengage and lack of accurate information on the profile, as detailed in Sect. 1. Finally, it has been shown that games can be an aid for gathering personal data, although this has not been unified within a chatbot yet to steer smoking trigger detection & smoker profiling.

3 Chatbot Design

Before diving into the design of the chatbot itself, first the application as a whole is presented. This contains the interface to converse with the chatbot, and the server which contains the chatbot, the database, and the smoking trigger detection module. Then the designed chatbot and games themselves and the smoking trigger detection module are described.

3.1 General Concept of the Application

The interface of the application is Facebook Messenger because it is the platform with the most users and the interface is familiar so it requires a lower threshold to start interacting with the chatbot[1]. A user can send messages to the chatbot through a Facebook page. The flow after the user sends a message until they receive an answer is given in Fig. 1: (1)-(8). The message first goes to the webhook (1) that is connected to the Facebook page who sends it to the server (2). The webhook endpoints at the server side give the message to the chatbot (3) who gets the necessary data to answer from the MongoDB database (4–5). After constructing its answer, the message is sent back with the post method at the server's side (6–7). The message goes back the same way (7–8).

The second flow in Fig. 1 (A)-(B) shows the loop that happens every fifteen minutes. The module trigger detection checks repeatedly if there is any new data in the database. If there is, the module either classifies it as a potential or

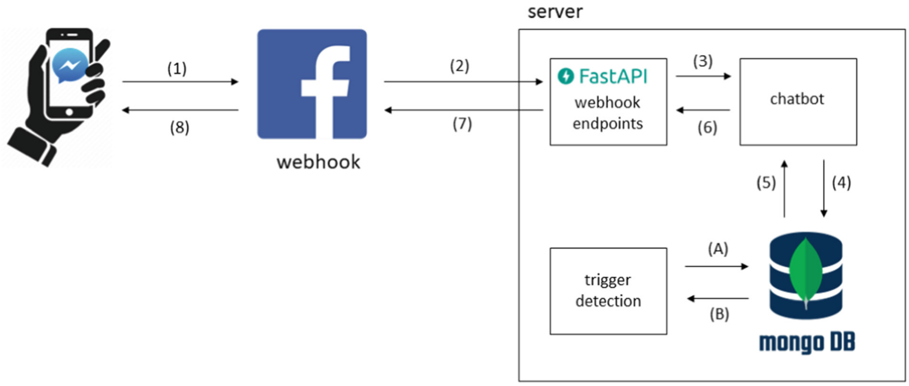


Fig. 1. Schema of the application

---

[1] https://datareportal.com/social-media-users.

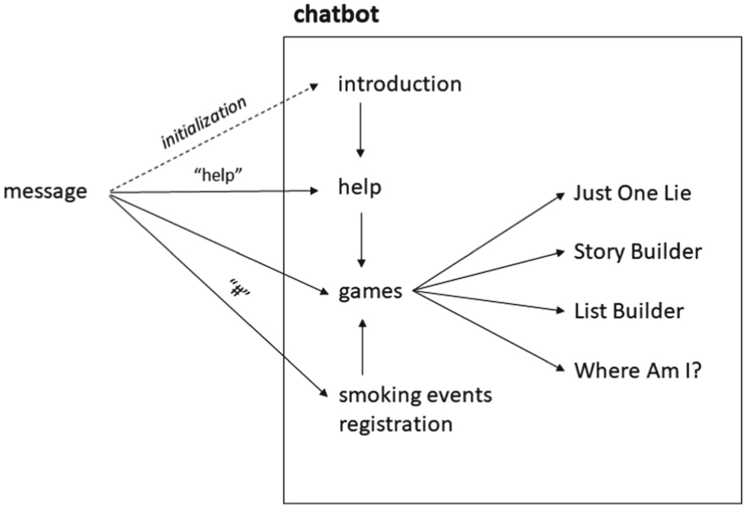


Fig. 2. Schema of chatbot component

detected smoking trigger. A potential smoking trigger is not confirmed yet, so the module stores the piece of data separately to process it further later.

Since there is a wide variety of smoking triggers, classifying them into categories makes processing a lot easier. These categories are based on previous smoking profiling and trigger studies [2,15] and are the following: negative emotions, stress, activities, smoking cues, substance abuse, location, moment of day/week, and general.

3.2 Chatbot

The chatbot exists out of different main logical components presented in Fig. 2 and divided according to the time of sending the message and the content. When starting the chatbot for the first time, it introduces itself and asks after the user's name. Then in the help part the user is informed about all the chatbot's functionalities and how to call upon help again. The main part exists out of the four different games: *Just One Lie*, *Story Builder*, *List Builder*, and *Where Am I*. These games were chosen based on the literature study of existing games that seemed to show most potential towards trigger detection. The last part deals with the smoking registrations and can be called upon by sending "#".

Every game has questions and subjects related to smoking. Only the combination of the output of all the games brings a total view of the user's smoking behaviour. A demo of the chatbot can be found here: http://predict.idlab.ugent.be/projects/imperio/.

Just One Lie. In this game one of the players gives two truths and one lie about a certain subject while the other guesses what the lie is. The smoking triggers that could be derived from this game involve hobbies, friends, family, stress factors, negative feelings, daily life, alcohol use, etc. Even smoking triggers themselves are acquired in this game. Moreover, the chatbot informs them about health effects and facts about smoking while dispelling myths around smoking.

Story Builder. One player starts with a prompt, and then one by one both the player and the chatbot each add a continuation to the story. This goes on until the players are satisfied with their story. The smoking triggers that could be derived from this game are the ones that happen in specific life events.

List Builder. One player starts with a word or group of words around a subject, and the other player answers with a word that begins with the last letter of the first player's word. This goes on until one of the two runs out of options. The smoking triggers that could be derived from this game are about hobbies, stress factors, and family. Even smoking triggers themselves are acquired in this game.

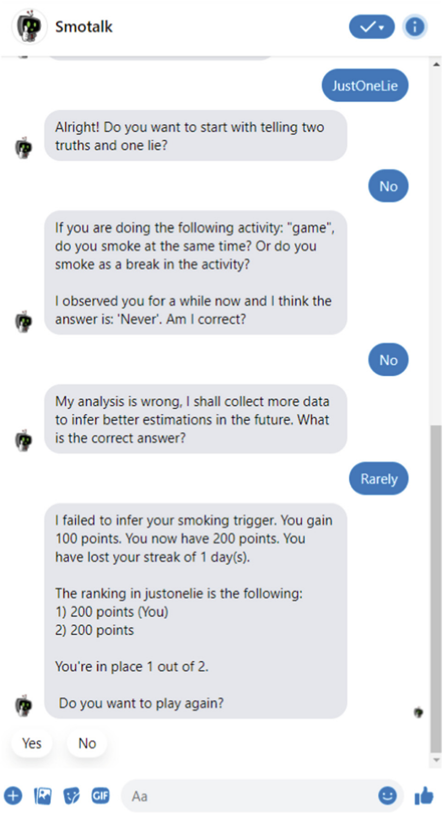
Where Am I?. One player is currently at a certain location and gives the other player hints about where they are. The other player guesses with every hint their location. The game ends when the other player guesses or gives up on guessing the place. In combination with smoking events, the smoking triggers that could be derived here are about a location.

The messages that the chatbot sends are static, and the different games have randomisation to keep the interest of the player. *Just One Lie* has 17 lies and 101 truths to educate the user. In the case of *List Builder* each of the four subjects has multiple options as answer for every letter of the alphabet. The hidden feature in *Story Builder* has for two of the smoking trigger categories and for all 6 genres a different story for the user to experience. In the game *Where Am I* the chatbot can guess 23 different places.

3.3 Smoking Trigger Detection

Processing of the data of the different games depends on the category and type of data. The different categories are: negative emotions, stress, activities (such as hobbies and possible breaks during these hobbies), smoking cues (such as seeing a cigarette, lighter, or other smokers), substance abuse (such as alcohol consumption), location, and moment of day/week. A last general category keeps track of all the smoking triggers that do not belong in the former categories. Emotion detection is done through a natural language processing model on data that is not yet classified into a category, such as a life event from *Story Builder*. If a negative emotion is detected in the life event, it is sorted into the category negative emotions. When the context is known and thus the category is known, the potential smoking trigger is integrated into a question to be evaluated later.

Evaluation of the potential smoking triggers is done in the hidden features in the games *Just One Lie* and *Story Builder*. The one in *Just One Lie* looks at



JustOneLie

Alright! Do you want to start with telling two truths and one lie?

No

If you are doing the following activity: "game", do you smoke at the same time? Or do you smoke as a break in the activity?

I observed you for a while now and I think the answer is: 'Never'. Am I correct?

No

My analysis is wrong, I shall collect more data to infer better estimations in the future. What is the correct answer?

Rarely

I failed to infer your smoking trigger. You gain 100 points. You now have 200 points. You have lost your streak of 1 day(s).

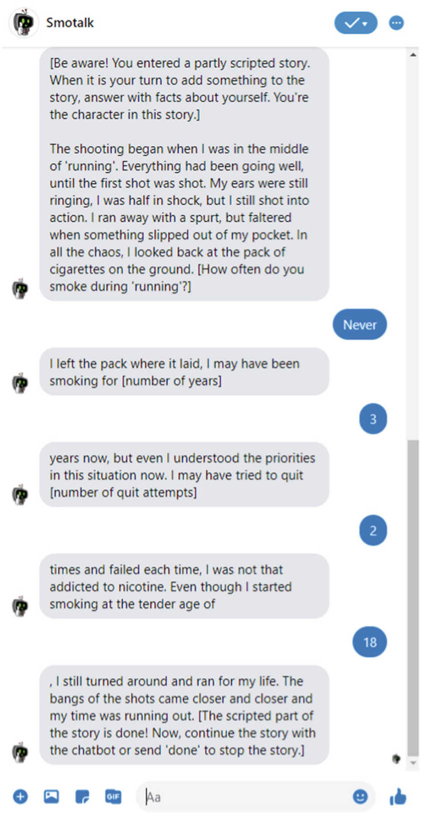
The ranking in justonelie is the following:
1) 200 points (You)
2) 200 points

You're in place 1 out of 2.

Do you want to play again?

Yes No

Fig. 3. Hidden Feature: Chatbot guesses severity of trigger [Just One Lie]



[Be aware! You entered a partly scripted story. When it is your turn to add something to the story, answer with facts about yourself. You're the character in this story.]

The shooting began when I was in the middle of 'running'. Everything had been going well, until the first shot was shot. My ears were still ringing, I was half in shock, but I still shot into action. I ran away with a spurt, but faltered when something slipped out of my pocket. In all the chaos, I looked back at the pack of cigarettes on the ground. [How often do you smoke during 'running'?]

Never

I left the pack where it laid, I may have been smoking for [number of years]

3

years now, but even I understood the priorities in this situation now. I may have tried to quit [number of quit attempts]

2

times and failed each time, I was not that addicted to nicotine. Even though I started smoking at the tender age of

18

, I still turned around and ran for my life. The bangs of the shots came closer and closer and my time was running out. [The scripted part of the story is done! Now, continue the story with the chatbot or send 'done' to stop the story.]

Fig. 4. Hidden Feature: Script requires severity of smoking trigger and personal data [Story Builder]

former smoking triggers in the same category and determines the likely severity of the new potential smoking trigger. The guess is either confirmed or refuted by the player. An example of this for the potential smoking trigger "game" is given in Fig. 3. The hidden feature in *Story Builder* acquires profile data through a scripted story that makes the player the main character. Moreover, a potential smoking trigger is integrated into each scripted story, and the player gives the severity when the potential smoking trigger appears in the story. An example of this with the potential smoking trigger "running" is given in Fig. 4.

3.4 Gamification

Gamification is integrated into every game through a point system, a streak feature and a leaderboard. Every game gives a minimum number of points, and more points can be gained the longer the game is played or the more effort is done. If a game is played multiple days in a row, a streak is formed. The longer the streak, the more points that can be accumulated every day. This streak feature encourages players to play the same game daily to gain more points. In order to give meaning to the point system, a leaderboard entices the player to beat other players. It speaks to their competitive side. This combination of gamification elements is implemented to motivate the player to play more games.

4 Study Design

The goal of the study is to evaluate the effectiveness of the chatbot to profile users and to study how engaging the users find the chatbot. The study started with 12 participants who fulfilled the following inclusion criteria:

- Age 18 and above
- A smoker (so someone who is an active smoker and is not currently trying to stop smoking)
- Has a Facebook account
- Is able to hold a conversation in English (since all the conversations with the chatbot are in English)

Even though the study is done in the domain of smoking cessation, the participants were not required to quit smoking since this was not necessarily needed for the detection of smoking triggers. This way the impact of the chatbot on the participants was limited, but still the required personal data could be acquired. The study was approved by the committee on ethics and data management of the faculty of Engineering and Architecture of Ghent University. All participants signed an informed consent form before the start of the study. The requirement for the study was to play at least one game with the chatbot daily and to register every smoking event with the chatbot. During the study there were 2 dropouts due to unforeseen circumstances (IDs 3 and 12) - they only participated for a week, but they did not ask to remove the data already acquired, so the already gathered data could still be used in the analysis.

The study lasted 15 d after which the participants filled in a questionnaire with questions about the gathered data, the chatbot, the smoking triggers, and general opinions.

5 Results

The first author K. Bosschaerts analyzed all the answers. It was a questionnaire specifically tailored to this study, but constructed in collaboration with experienced behavioral change user researcher (and co-author) J. Stragier. The used questionnaire and the code base are made available at: https://github.com/predict-idlab/chatbot-smoking-profiling.

Engagement is an important factor for effective behavioral change, so this is discussed first. Then the participants' answers from playing the games is looked at next. At last, the user experience gained from the questionnaire is given.

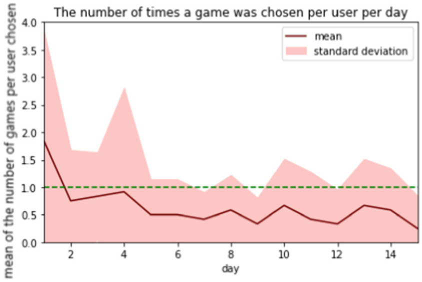


Fig. 5. The average number of times a game was chosen per user per day (number of participants shown (n) = 10)

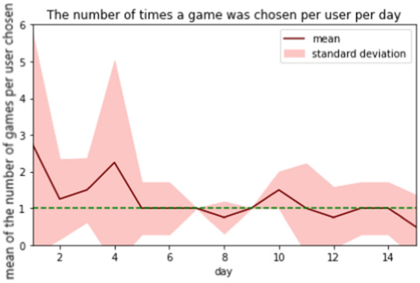


Fig. 6. The average number of times a game was chosen per user per day (selection of 4 participants that engaged every day)

5.1 Engagement

The requirements stated that participants needed to play at least one game per day and that they had to register every smoking trigger, but only a few participants abode by it. The mean reason for this lower engagement is that the participants were not required to stop smoking, so they had less motivation to learn about their smoking behavior. Figure 5 shows that the mean number of times a game was chosen per user per day lays below the minimum of one - the green line. A selection of four participants did fulfill this requirement as shown in Fig. 6. Nevertheless, the other participants' data can also be used to measure some aspects of engagement, and their answers can still be used to evaluate the games.

The participants who did not play the game daily, were also the ones that stopped registering their smoking triggers midway. For others, a decline in smoking registrations can be seen in Fig. 7. Here the drop-outs - the two participants who quit in the middle of the study - were removed. Manually registering the smoking events is tedious, so naturally, participants register less and less smoking events as the study goes on. The impact of this decline is limited to two categories: moment of day/week and locations. Both of these categories rely on accurate data to infer the smoking triggers, so results here are skewed.

A chatbot has to have a Conversation-turns Per Session (CPS) - the average number of conversation-turns between the user and the chatbot in a conversational session - of at least 10 to qualify as a social chatbot [16]. The CPS of all participants comes down to 10 with a standard deviation of 9.18, while the CPS of the selection of 4 participants is 10.29 with a standard deviation of 6.96. Thus, in both cases the chatbot's engagement qualifies it as a social chatbot. The standard deviation of the CPS of all participants lies higher than the selection because some non-daily players had very long sessions with the chatbot.

The leaderboard in Table 1 gives an other view of game engagement. The user IDs in bold are the four participants that engaged daily. A few participants mentioned that the leaderboard spoke to their competitive side and made them

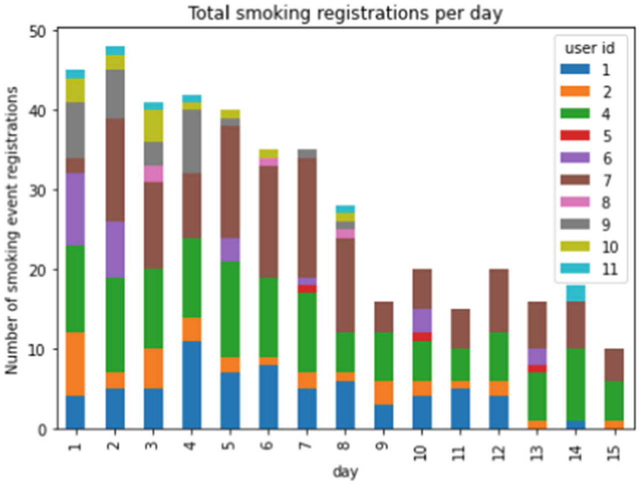


Fig. 7. Total smoking registrations per day (n=10)

want to get to the top. The game *Just One Lie* was the most popular one as it is short and it has additional educational value about smoking. The games *Where Am I* and *Story Builder* were preferred by different participants, but *List Builder* was overall the least favourite.

Table 1. Leaderboard; The rank represents all the players from most played to least played, with their ID in the second column. The third column gives the total amount of points per player over all the games, while the last four columns gives the total amount of points per game per player.

Leaderboard						
rank	id	total points	Just One Lie	Where Am I	List Builder	Story Builder
1	7	4480	0	4480	0	0
2	**1**	3430	930	500	1070	930
3	**4**	2850	1780	570	410	90
4	6	2810	2520	0	120	170
5	**2**	2540	2330	70	70	70
6	3	1090	500	260	210	120
7	**5**	1030	710	320	0	0
8	12	860	630	230	0	0
9	11	700	520	0	180	0
10	8	510	450	0	0	60
11	9	30	0	0	30	0
12	10	0	0	0	0	0

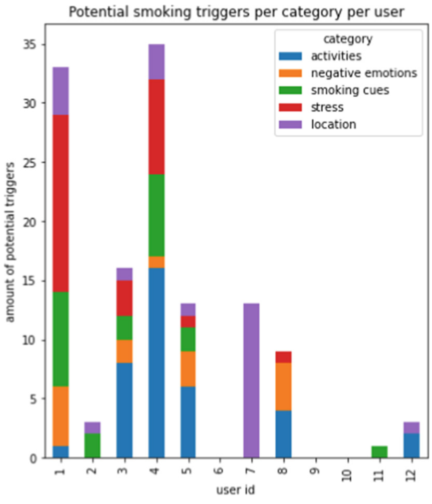


Fig. 8. Potential smoking triggers per category per user (n=12)

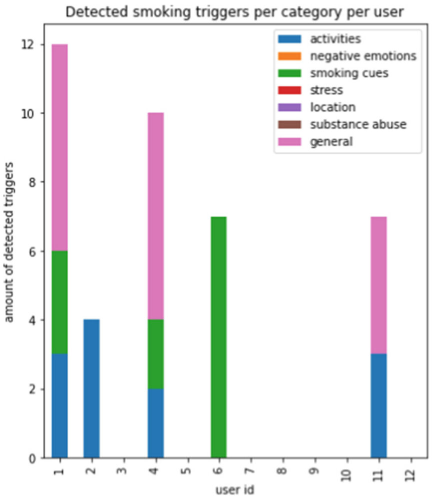


Fig. 9. Detected smoking triggers per category per user (n=12)

5.2 Analysis of Participants' Answers

The quality of the data of both *Just One Lie* and *Story Builder* cannot be sufficiently used for smoking profiling and trigger detection. Due to the nature of the games, the data is riddled with useless information, mistakes, and bad answers. *List Builder* and *Where Am I* are a lot better in comparison. *List Builder* brings high quality, short data. Even though *Where Am I*'s locations were sometimes too specific and in the format of a sentence, most of the data is clean and useful.

A lot of potential smoking triggers, i.e. triggers that could be possible smoking triggers for a certain smoker, but that are not confirmed or refuted yet by the participants during the study, were found as shown in Fig. 8. However, very few of the potential triggers could be confirmed or refuted by the chatbot during follow-up games. The few detected smoking triggers through the games are illustrated in Fig. 9. This was a consequence of the small number of hidden feature games that were played. Half of the participants indicated in the questionnaire that they were not aware of any smoking trigger they had, while the other half knew about one, two, or three triggers. Smokers are often not aware of their own smoking triggers, and the chatbot is able to detect these for them. None of the substance abuse triggers were found, even though a few users indicated that alcohol was a trigger for them. More games focusing on this category could prevent this problem. The location category required the user to play the game *Where Am I* multiple times at the same location. Hardly anyone played the game more than twice on the same location, however, but combining the game *Where Am I* with Global Positioning System (GPS) data solves the issue. Since the category moment of day/week was calculated in the end, every user gained a detected smoking trigger in that category. Because of the unreliable manual registrations of the smoking events, as shown in Fig. 7, these are not accurate though. The selection of categories was a good choice in the participants' opinion.

5.3 User Experience

The opinion of the participants about the chatbot in Fig. 10 is very black-and-white, with persons either liking the games, or not liking them at all. Participants who do not like games in general, as indicated by them in the closing questionnaire, did not like the method used and they also are causing most of the lower ratings in Fig. 11. They thought it was taxing to talk with the chatbot every day, as can be seen in Fig. 12, and would not like to use the chatbot outside the study. The group who does like games or is neutral towards them, is more divided in their opinions. One common thing that all users experience is that talking with the chatbot takes time and is hard to fit into their daily life. With IMPERIO we will try in the future to determine the optimal intervention point to resolve this issue. The users will get a notification and they will be more inclined to interact with the chatbot.

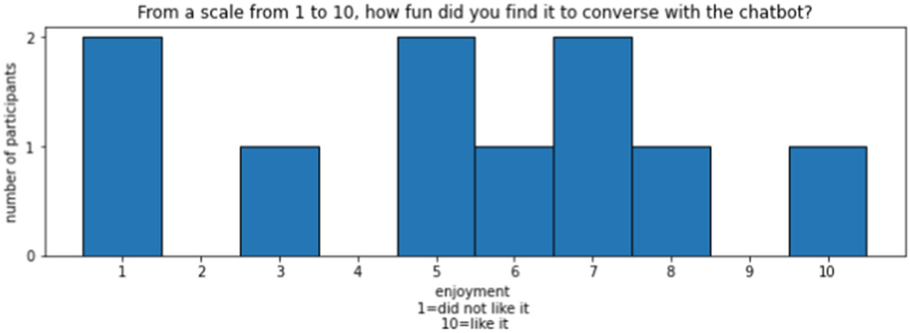


Fig. 10. User enjoyment of the chatbot (n=10)

Even though there are a lot of participants who did not like this gathering of their data through games, when they were asked if they would prefer a questionnaire instead, only three participants were more inclined towards the questionnaire, as shown in Fig. 13. This means that there is potential in the method of the study, but that the games need be refined some more. A few adjustments in the implementation have to be made before it can be put to use for a broader public. If some further testing needs to be done, maybe an additional inclusion criteria for the participants has to be that they have to like games or minigames.

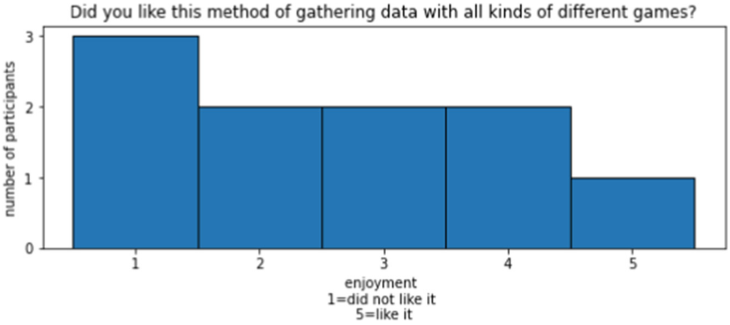


Fig. 11. User enjoyment of the method of gathering data (n=10)

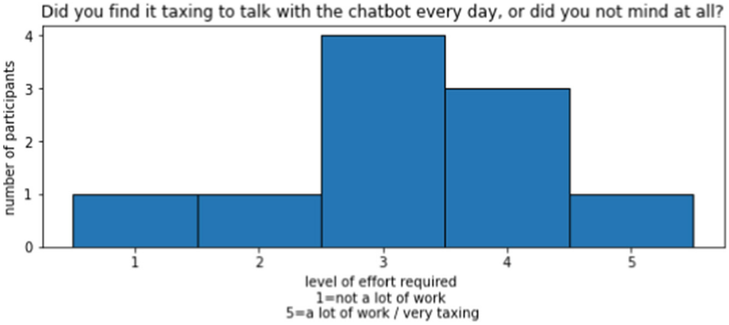


Fig. 12. Level of effort required to chat with the chatbot (n=10)

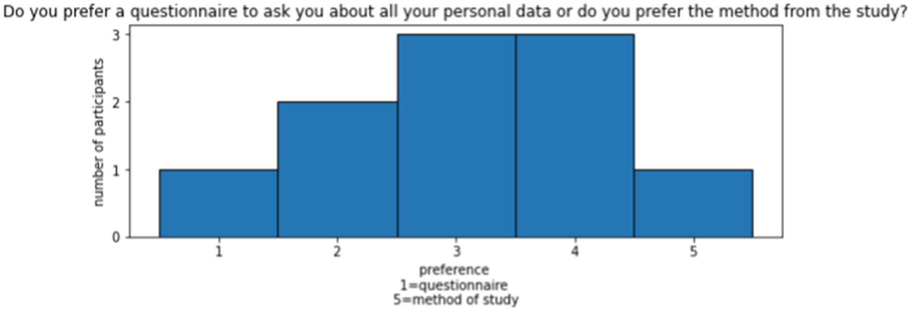


Fig. 13. Preference between questionnaire or study method to ask about personal information (n=10)

6 Conclusion

When personalisation is needed in an application, a questionnaire is mostly used. Such a questionnaire is tedious to fill in, and consequently often is not completed. In an attempt to look for an alternative, the first few steps were taken in this paper. This approach of combining a chatbot and games to engage users in providing as much personal data as possible while keeping their interest in doing so, can be a possible alternative for a questionnaire after more research around this topic is done.

The chatbot was tested by 12 participants during a period of 15 d. Since the user sample is rather limited, the conclusions written here are only an indication. The engagement with the chatbot was good since it achieved the minimal 10 CPS needed for a social chatbot. Although the chatbot was not able to convince the people who did not like games of its charms, some of these people still recognized that they preferred it above a boring questionnaire. The hurdle of filling out a questionnaire is thus greater than the participants' disinterest into games. For the people who do like games, this project was found to be a great alternative for a questionnaire as soon as the problems with the current chatbot are fixed, i.e. poor data quality in some games, more focus on underrepresented categories of triggers, and changing the manual smoking event registration to an automatic solution.

Acknowledgement. This work was partly funded by the Fond Wetenschappelijk Onderzoek (FWO) project IMPERIO (G0D5322N): http://predict.idlab.ugent.be/projects/imperio/. B. Steenwinckel is funded by a strategic base research grant of FWO (1SA0219N).

References

1. Almusharraf, F., Rose, J., Selby, P.: Engaging unmotivated smokers to move toward quitting: design of motivational interviewing-based chatbot through iterative interactions. J. Med. Internet Res. **22**(11), e20251 (2020)
2. Brown, A.E., Carpenter, M.J., Sutfin, E.L.: Occasional smoking in college: who, what, when and why? Addict. Behav. **36**(12), 1199–1204 (2011)
3. Daley, K., Hungerbühler, I., Cavanagh, K., Claro, H., Swinton, P., Kapps, M.: Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. Front. Digit. Health **2**, 576361 (2020)
4. Derksen, M.E., van Strijp, S., Kunst, A.E., Daams, J.G., Jaspers, M.W.M., Fransen, M.P.: Serious games for smoking prevention and cessation: a systematic review of game elements and game effects. J. AMIA **27**(5), 818–833 (2020)
5. Dosovitsky, G., Pineda, B.S., Jacobson, N.C., Chang, C., Escoredo, M., Bunge, E.L.: Artificial intelligence chatbot for depression: descriptive study of usage. JMIR Form. Res. **4**(11), e17065 (2020)
6. Fadhil, A., Gabrielli, S.: Addressing challenges in promoting healthy lifestyles: the al-chatbot approach. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, pp. 261–265. PervasiveHealth, New York, NY, USA (2017)
7. Hébert, E., et al.: A mobile just-in-time adaptive intervention for smoking cessation: pilot randomized controlled trial. J. Med. Internet Res. **22**, e16907 (2020)
8. Hébert, E.T., et al.: An ecological momentary intervention for smoking cessation: the associations of just-in-time, tailored messages with lapse risk factors. Addict. Behav. **78**, 30–35 (2018)

9. Paay, J., Kjeldskov, J., Skov, M.B., Lichon, L., Rasmussen, S.: Understanding individual differences for tailored smoking cessation apps. In: Proceedings of the 33rd Annual CHI, pp. 1699—1708 (2015)
10. Perski, O., Crane, D., Beard, E., Brown, J.: Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. Digit. Health **5**, 2055207619880676 (2019)
11. Rajani, N.B., Mastellos, N., Filippidis, F.T.: Impact of gamification on the self-efficacy and motivation to quit of smokers: observational study of two gamified smoking cessation mobile apps. JMIR Serious Games **9**(2), e27290 (2021)
12. Rolstad, S., Adler, J., Rydén, A.: Response burden and questionnaire length: is shorter better? A review and meta-analysis. Value Health **14**(8), 1101–1108 (2011)
13. Ryan, K., Dockray, S., Linehan, C.: A systematic review of tailored eHealth interventions for weight loss. Digit. Health **5**, 1–23 (2019)
14. Sazonov, E., Lopez-Meyer, P., Tiffany, S.: A wearable sensor system for monitoring cigarette smoking. J. Stud. Alcohol Drugs **74**, 956–64 (2013)
15. Shiffman, S., et al.: Smoking patterns and stimulus control in intermittent and daily smokers. PLoS ONE **9**, 1–14 (2014)
16. Shum, H.Y., He, X., Li, D.: From Eliza to Xiaoice: challenges and opportunities with social chatbots (2018)
17. Spears, C.A., et al.: Text messaging to enhance mindfulness-based smoking cessation treatment: program development through qualitative research. JMIR Mhealth Uhealth **7**(1), e11246 (2019)
18. Stichting tegen Kanker: Rookenquete (2019). https://www.kanker.be/sites/default/files/stichting_tegen_kanker_-_rookenquete_2019_-_def.pdf
19. Sutfin, E.L., et al.: First tobacco product tried: associations with smoking status and demographics among college students. Addict. Behav. **51**, 152–157 (2015)
20. Vogel, E.A., Humfleet, G.L., Meacham, M., Prochaska, J.J., Ramo, D.E.: Sexual and gender minority young adults' smoking characteristics: assessing differences by sexual orientation and gender identity. Addict. Behav. **95**, 98–102 (2019)
21. World Health Organisation (WHO): Noncommunicable diseases (2018). https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases
22. Zhao, J., Freeman, B., Li, M.: Can mobile phone apps influence people's health behavior change? An evidence review. J. Med. Internet Res. **18**(11), e287 (2016)

Specification of Quality of Context Requirements for Digital Phenotyping Applications

Luís Eduardo Costa Laurindo(✉), Ivan Rodrigues de Moura, Luciano Reis Coutinho, and Francisco José da Silva e Silva

Laboratory of Intelligent Distributed Systems (LSDi),
Federal University of Maranhão (UFMA), 65080-805 São Luís, Brazil
{luis.laurindo,ivan.rodrigues,fssilva}@lsdi.ufma.br, luciano.rc@ufma.br

Abstract. Digital phenotyping applications use sensor data from personal digital devices (e.g., smartphones, smart bands) to quantify moment-to-moment human phenotype at the individual in-situ level. Ensuring the quality and distribution of the data used is essential requirement in the domain of these applications. Context Quality (QoC) refers to the Information Quality (QoI) used and the Quality of Service (QoS) level of information distribution. QoI is measured by parameters that define how reliable the information is. On the other hand, QoS is provided by specifying the quality of service for distributing context data. Some aspects can degrade the QoC of the application, such as information from sensors being imprecise, wireless communication technologies used in the acquisition and distribution of information, scalability problems can cause information delay, and intermittent connection due to user mobility can result in data loss. Therefore, this study conceives a process for incorporating QoC requirements and a Domain-Specific Language (DSL) to specify these requirements in digital phenotyping applications. A case study was carried out where the scenario of an application for monitoring workers' health was considered. It was possible to prove the expressiveness and simplicity of the proposed language when using it to define the instances of the application classes responsible for the acquisition and distribution of context information.

Keywords: Digital Phenotyping · Acquisition and Distribution · Quality of Context (QoC) · Incorporation of QoC Requirements · Domain-Specific Language

1 Introduction

Smartphones and wearable devices are part of people's daily lives and have sensors that capture the user's context, and environment [13,18]. Computational methods can use this context information to make inferences about social, behavioral, and cognitive aspects of individuals [12,14]. For example, it is possible

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 651–665, 2023.
https://doi.org/10.1007/978-3-031-34586-9_43

to identify if the user is performing some physical activity using location and acceleration data or if he is carrying out a conversation based on the data captured by the smartphone's microphone.

Digital Phenotyping is a research area that aims to use context data from mobile devices to infer the individual's health status [19]. Research in digital phenotyping focuses on developing new solutions to complement and extend traditional sources of clinical data. Digital phenotyping solutions need to ensure an acceptable level of data quality to obtain greater decision-making accuracy due to their applicability in the healthcare field. Quality of Context (QoC) refers to the Quality of Information (QoI) used as context information and the Quality of Service (QoS) for distributing this information [2]. With the resources offered by QoC, it is possible to establish quality contracts between service providers and consumers, select better context sources, increase application efficiency by optimizing energy and bandwidth consumption by adapting the frequency of information dissemination and improving the quality of the user experience [5].

In the literature, one can find several parameters related to QoC to specify the application quality requirements [8,10]. For example, in an application that makes decisions in near real-time, the age parameter is crucial, as outdated information may not represent the individual's current context. Furthermore, by distributing this information through the smartphone with battery limitations and an increase in energy consumption when sending data over the network, it is soon possible to specify the desired frequency for distributing this information.

Digital phenotyping applications run in environments that can degrade QoC. First, they use sensor data which can generate inaccurate and erroneous readings [6]. Second, they deal with intermittent connection by allowing individuals to move and situations where the current network does not meet application requirements, resulting in data loss. Finally, a digital phenotyping infrastructure allows the monitoring a set of individuals, and problems related to scalability may occur. Therefore, it is necessary to specify QoC parameters to meet the requirements of each application in order to evaluate the context information used in them. In addition to evaluating the information, it is necessary to evaluate the quality of the distribution service. It is also critical to monitor compliance with these requirements to resolve potential issues that degrade the QoC of the application.

When designing these applications, it is necessary to develop or use software platforms to perform the acquisition, inference, and distribution of context information. Digital phenotyping is currently a very active research area [11,12]. Regarding middleware platforms and several application proposals were presented, such as [1,7,15,19,20]. However, the treatment of QoC in the field of digital phenotyping is still incipient in the literature, especially in middleware platforms focused on this area. Therefore, when considering the importance and requirements of these applications, this study conceives an approach that has a process for incorporating QoC requirements and a Domain-Specific Language (DSL) for specifying these requirements in digital phenotyping applications.

The article is divided as follows: Sect. 2 presents the theoretical foundation; the 3 section exposes related works; Sect. 4 presents the proposed solution;

Sect. 5 presents a case study; and finally, Sect. 6 presents the conclusions and future works.

2 Background

2.1 Digital Phenotyping

Torous et al. [19] define Digital Phenotyping as the "moment-by-moment quantification of human phenotype at the individual level in-situ using data from smartphones and other personal digital devices." Personal digital devices are present in the individual's daily life (for example, smartphones, smartbands) that collect a set of data on behavioral and health aspects useful in the domain of these [14] applications.

According to Mendes et al. [11] the digital phenotyping process (illustrated in Fig. 1) begins with the collection of raw data from sensors, whether physical (e.g., GPS, accelerometer, heart rate) or virtual (e.g., phone calls, screen time on, apps used). After collecting the raw data, behavioral and health events are inferred. Behavioral events represent actions performed by the individual (e.g., the time interval in which he socialized with the family). Health events represent the individual's physiological state based on vital signs (e.g., heart rate, blood pressure, blood oxygenation). After the inference of these events, behavioral patterns are obtained (e.g., the pattern of mobility, sociability, and physical activity). These patterns refer to routine situations of the individuals (e.g., the individual always sleeps, from Monday to Friday, at 23:00). Finally, these behavioral patterns are used in prediction and diagnostic applications.

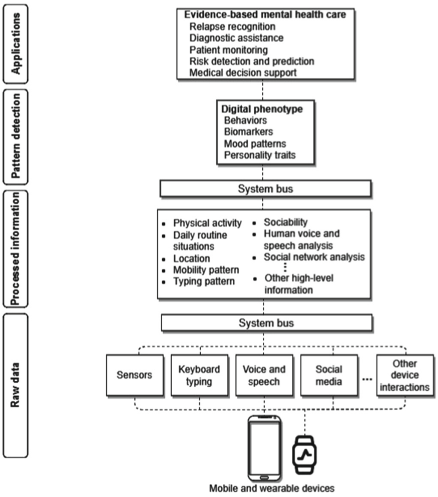


Fig. 1. The process of digital phenotyping [11].

The development of digital phenotyping applications requires components responsible for collecting context data from sensors embedded/connected to personal digital devices. In addition, developers must implement software components to infer behavioral events and distribute collected and inferred data to external servers via wireless communication infrastructure. Finally, on the server, it is necessary to implement components for managing large volumes of data and machine learning models to infer behavioral patterns related to health aspects. [11].

2.2 Quality of Context

Context is the information used to characterize the situation of an entity (e.g., people, objects, or places) that influence an agent's decisions [4]. Below are described just some of the various parameters present in the literature used to quantify the degree of information quality, in addition to allowing the definition of the quality of the distribution service [8,10].

The **Accuracy** represents an estimate of how close the context information is to the actual value. The device manufacturer normally provides the accuracy value when the information comes from a physical sensor. The **Confidence** estimates the degree of certainty of the information provided by the information source. The **Measurement Interval** indicates the time interval between successive readings. The **Delay** represents the time elapsed between sending the message by the producer and its arrival at the consumer. Even though it is a QoS parameter, it is possible to insert the computed time in the context data as a meta-information related to QoI. **Completeness** indicates how complete is the context information received by a consumer. It can be calculated based on the ratio between the sum of the weights of the available attributes and the sum of the weights of the attributes required by the consumer. The **Validity Time** indicates the validity of the information. The information producer can specify the validity of the information as he has produced it. However, the consumer can decide whether or not to accept the shelf life specified by the producer. The **Age** can also be used to check the validity time of the information. It indicates the difference between the current instant and the information measurement time. Finally, the **Total Delivery Time** indicates the time from the measurement of the information to its delivery to the consumer.

To define the quality level of the information distribution service, we have the **Reliability** that determines whether the distribution service should adopt the best effort policy (best-effort), when there is no guarantee of delivery of information, or use delivery and retransmission if the context information is not delivered to the recipient. The **Refresh Rate** allows the consumer to define how often context information should be received regardless of how often the data is produced. The **Delivery Time** indicates the maximum time a consumer is willing to wait for information. The **Latency Control** defines an additional delay to the producer and consumer of the information. By setting a delay, messages are grouped into a queue and sent or received in a single burst. The **History** allows the consumer to store the information for some time feasible for him. By using this parameter **Order of Destination**, the messages stored in the History can be organized according to their timestamp of publishing or receiving. The consumer can still specify the validity time of the information with the parameter **Lifetime**. Through this parameter, the consumer defines when messages should be removed from the History. **Retention** allows the information producer to indicate that he wants to retain the last message sent to new consumers as they arise. Finally, **Liveness** allows producers to send consumers the status of their service, indicating whether they are still active. The consumer must indicate whether he wants to receive this alert.

3 Related Work

We used as related works the studies selected in the Systematic Literature Review (RSL) conducted by Mendes et al. [11]. The study authors presented software platforms designed to support digital phenotyping studies. Based on this review, we present in this section works that conceive software platforms to perform the acquisition and distribution of context information.

AWARE [7] is a reusable Android software platform focusing primarily on the acquisition, distribution, and context inference. The platform provides a library that allows the developer to design their application. Its architecture is composed of the Aware Client and Aware Server layers. The Aware Client layer is responsible for acquiring context information from physical and virtual sensors. After the acquisition, information is processed, and high-level situations are inferred. Processing is carried out through plugins. Each plugin is responsible for some situation of interest (e.g., sociability, mobility, physical activity, sleep). The Aware Server layer has a cloud database and dashboard capabilities for information visualization. The information presented in real time is sent via the Message Queuing Telemetry Transport (MQTT)[1] protocol. The study addresses some measures to minimize data loss. The first is to use MQTT's QoS requirements, and the second is to prevent data collection processes from being interrupted by Android.

SituMan [17] is a reusable software platform that identifies situations in the individual's routine based on data from smartphone sensors. Situation inference is used to solicit self-reports at opportune times. It allows collection related to the location and activity that the individual is performing. The authors recognize that the sensor data can sometimes generate inaccurate data, as in the case of the GPS used by the authors in the study, but the solution does not address the provision of QoC mechanisms. When using inaccurate location information, it may not represent the exact location where the individual is.

Beiwe [19] is a platform that collects raw data from sensors and smartphone usage aspects. Two main components make up Beiwe: a web app and a mobile app. The web application allows researchers to specify the application's content, the sensors used in data collection, and its refresh rate. The mobile application is responsible for acquiring and distributing context information. They are stored in

---

[1] https://mqtt.org/.

a buffer to be sent later. Buffering information is one of the important QoS requirements, as, at certain times, the application may not have an internet connection, so this information is stored to be sent later when establishing the connection.

Purple Robot [15] is a framework that supports the creation of mobile applications which collect data from sensors through an authoring tool web that helps researchers who do not know software to develop their applications to acquire data from smartphone sensors to carry out their studies.

Funf [1] provides a set of functionality that allows the collection and distribution of context information. Funf has a Funf Manager component responsible for selecting the sensors used to collect context information and configuring the data collection frequency to minimize battery consumption. The possibility of changing the data sampling rate is relevant when considering device battery limitations and application bandwidth consumption. The solution also has a buffer where data is temporarily stored. The Funf sent this data every three hours to a server in the cloud. Being able to temporarily store information in a buffer is also a QoS requirement, as constantly accessing the network can consume increasing device power consumption.

Finally, Sensus is a tool that consists of two [20] mobile applications. The first one is responsible for collecting data from sensors from the monitored individual's smartphone. A server processes the collected data in the cloud. The other application is used by the healthcare professional to manage the study. The study is managed by a protocol designed by the professional. The protocol contains the information necessary to apply the study and the sensor data collected from the subject's smartphone.

3.1 Considerations

Analyzing the related works, we identified that the platforms designed to acquire and distribute context information within the scope of digital phenotyping do not broadly address the QoC requirements necessary for these applications. However, the treatment of QoC in the domain of digital phenotyping is still incipient in the literature, particularly in middleware platforms focused on this area.

Each application may require different levels of QoC. Information that does not meet the requirements required by the application may be useless for the context. Therefore, this study contributes to the literature in proposing mechanisms that facilitate the specification of QoC requirements in digital phenotyping applications, considering both QoI and QoS aspects.

4 Proposed Solution

4.1 Process

The proposed solution has a process for incorporating QoC requirements in developing these applications and monitoring their execution. Figure 2 presents the steps of this process. The process consists of five steps, namely: specification, transformation, implementation, evaluation and monitoring, and finally, visualization.

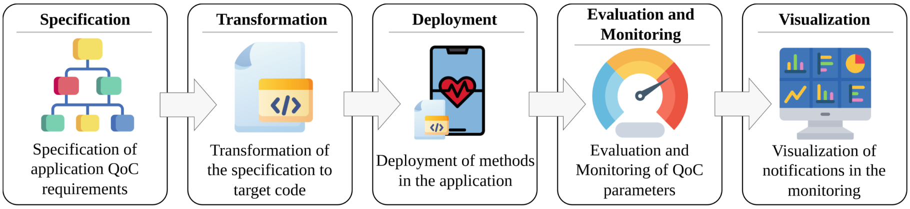

Specification
Specification of application QoC requirements
Transformation
Transformation of the specification to target code
Deployment
Deployment of methods in the application
Evaluation and Monitoring
Evaluation and Monitoring of QoC parameters
Visualization
Visualization of notifications in the monitoring

Fig. 2. Process for incorporating QoC requirements into digital phenotyping applications.

The first step is to specify the application's QoC requirements with the help of a domain language. A DSL is conceptualized as a set of models, defined by a metamodel, that corresponds with an abstract syntax and is represented by one or more concrete syntaxes [9]. After formalizing the requirements, the transformation to the target code occurs automatically through a conversion process. After the transformation stage, the developer, in possession of the artifact generated based on the DSL, can use it in the design of his application by importing the generated source code into his project. Next, in the evaluation and monitoring stage, the context information is selected based on the specified QoC parameters, and event logs are generated regarding fulfilling the requirements. Finally, the visualization step involves projecting these data events onto the Dashboard tool.

4.2 Proposed Metamodel

For the design of the DSL, a problem domain analysis was first performed to identify the concepts, abstractions, and relationships between the entities. This domain analysis produced an abstract syntax that corresponds to a metamodel. The proposed metamodel was architected based on the Eclipse Modeling Framework[2] (EMF) pattern. EMF consists of a modeling framework based on a data model with code generation capabilities. Figure 3 presents the abstract syntax of the proposed DSL. She defines all identified concepts and their respective relationships. The concepts must be described to be easily understood by the user. The abstract syntax also includes structural metamodel semantics to define the rules and constraints of relationships between domain classes.

In digital phenotyping, applications consume context information from, for example, sensors. In addition, these applications distribute this information to other applications. Consumption and dissemination of context information are defined in the metamodel as services. Each service is associated with one or more context information. Each context information comprises a specific type of data (e.g., heart rate), a unit of measurement (e.g., bpm), and the source precision value.

---

[2] https://www.eclipse.org/modeling/emf/.

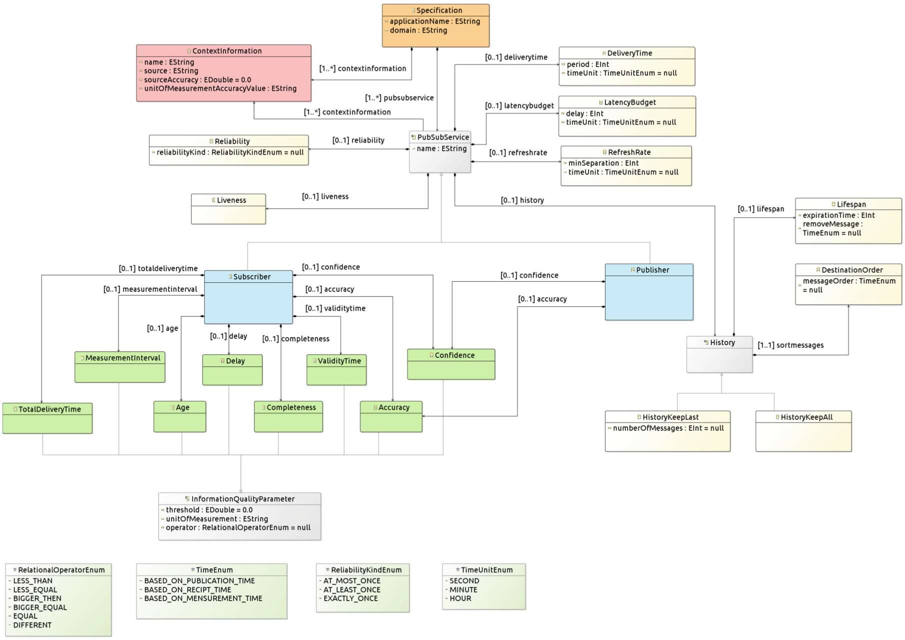


Fig. 3. Proposed metamodel.

For each service, it is possible to associate different QoC parameters. The QoI parameters, highlighted in green, are specified to ensure that the information received or sent will have a certain degree of quality required by the application. When defining a QoI parameter, it is necessary to define a threshold value, its equivalent unit of measure, and a relational operator. The relational operator is used to compare the value of the QoI parameter contained in the information with the specified value. For this, the context information must be annotated with the QoI meta-information.

The QoS parameters, highlighted in yellow, are specified to guarantee the distribution service quality. For example, it is possible to set the reliability level of data delivery by specifying the Reliability parameter and its type. By setting it to At Least Once, the service disseminates the information and will receive a puback to confirm receipt. For the Exactly Once type, a handshake is performed. This action is required to confirm the delivery of the information. If there is no such confirmation, the context information is retransmitted. It is also possible to set the QoS level when consuming the context information. For example, we can set the service that receives data from the heart rate sensor to have a Refresh Rate of 1 s.

4.3 Transformation and Incorporation of QoC Requirements

To incorporate the specified QoC requirements, it is necessary to use Middleware platforms with the mechanisms to implement the requirements in the proposed metamodel. Therefore, the process of transforming the model into Middleware-specific source code begins with the creation of code generator modules. Figure 4 presents the process necessary to carry out the transformation of the specification into the target code.

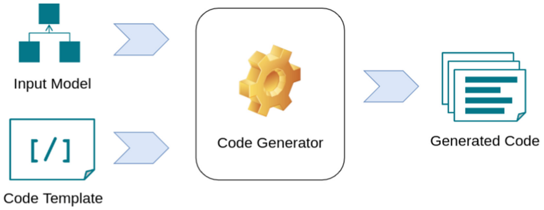


Fig. 4. Target code generation process for Middleware.

Implementing the MOF Model to Text Language (MTL) standard can be used to transform an EMF model into target code. The Acceleo tool implements the MTL standard and comprises two main structures: models and queries. Templates have a set of Acceleo instructions for generating text. The queries are performed using the Acceleo Query Language (AQL) and are responsible for extracting information from the EMF specification. After generating the destination codes referring to the services that consume (Subscribers) and distribute (Publishers) context information, the developer can incorporate them into the application.

5 Case Study

5.1 Worker Health Monitoring Mobile System

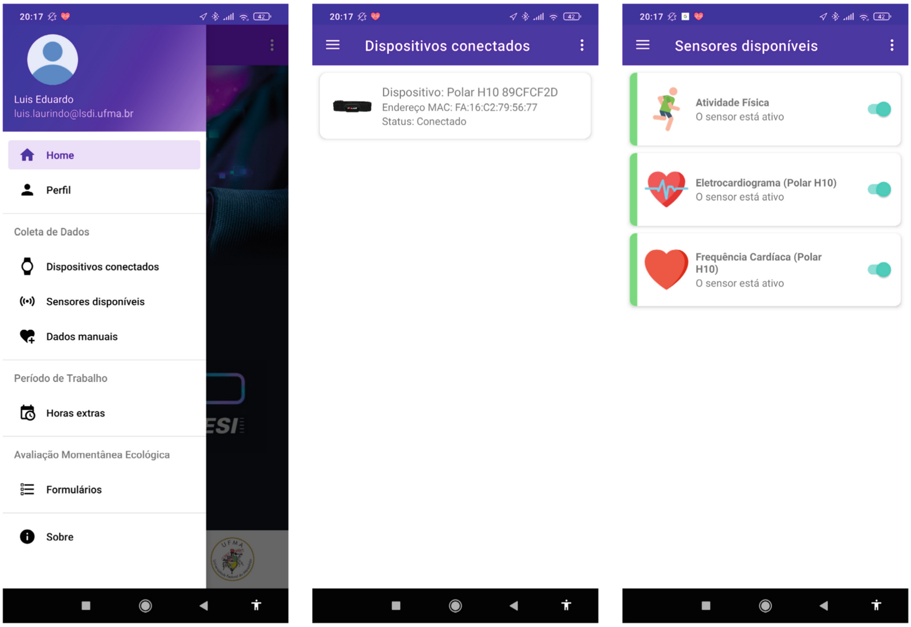
Mobile health monitoring systems are applications that run on mobile devices and aim to infer users' health status based on information derived primarily from sensors. These applications use wireless communication technologies to acquire data from wearable devices of monitored users and disseminate this information. As a case study, we implemented a mobile app that monitors workers' vital signs and physical activity during working hours.

We designed a worker health monitoring app to connect devices with Bluetooth communication technology. In addition to performing the collection, the application infers situations about the worker's health status and disseminates this information through the MQTT protocol to consumer applications. One is a SpringBoot[3] application that runs on a cloud server. It is responsible for

---

[3] https://spring.io/projects/spring-boot.

receiving the data by storing it. Another application consists of a Dashboard that presents to the health professional at a time close to the worker's current state of health. The Dashboard also allows the professional to consult historical data. Figure 5(a) shows the Polar H10 device connected to the mobile application. Figure 5(b) shows that the mobile application collects data from heart rate, electrocardiogram, and activity sensors. Finally, Fig. 6 shows the Dashboard presenting the data in real-time.



Dispositivo: Polar H10 89CFCF2D
Endereço MAC: FA:16:C2:79:56:77
Status: Conectado

Atividade Física
O sensor está ativo

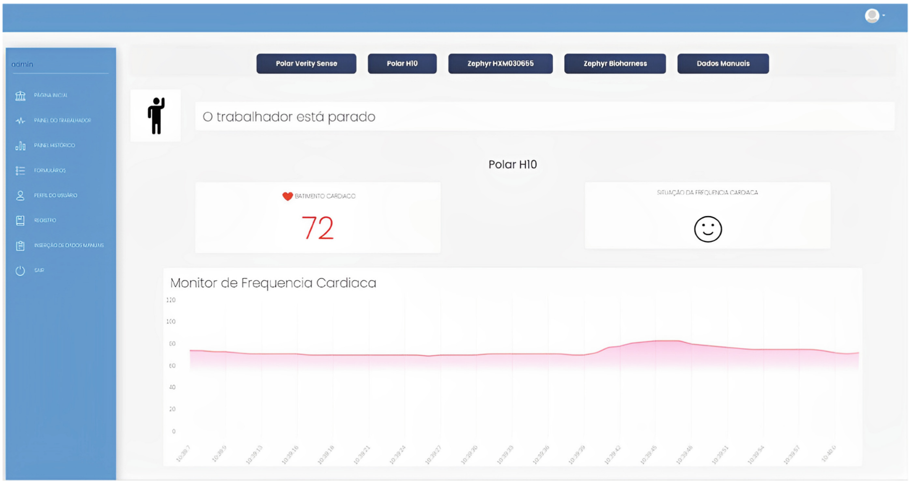
Eletrocardiograma (Polar H10)
O sensor está ativo

Frequência Cardíaca (Polar H10)
O sensor está ativo

(a) (b) (c)

Fig. 5. (a) Screen that shows options menu. (b) Screen that shows connected devices. (c) Screen showing available and active sensors.

The CDAL/CDDL Middleware platform was used for data acquisition and distribution. It is an extension of the M-Hub/CDDL Middleware [8,16], designed to facilitate the development of Internet of Things (IoT) applications with QoC requirements. Middleware is composed of two layers, they are: the Context Data Acquisition Layer (CDAL) and the Context Data Distribution Layer (CDDL). The CDAL layer runs on Android mobile devices to acquire context data from Smartphone and smart personal devices that have technologies such as Bluetooth Classic (BT) and Bluetooth Low Energy (BLE). On the other hand, the CDDL is responsible for processing and distributing the context data obtained through the CDAL and can be executed on Android devices, desktop applications and cloud servers. It provides developers of mobile applications, which use data from



O trabalhador está parado

Fig. 6. Presentation of near real-time data through the dashboard.

sensors from smartphones and wearable devices, mechanisms to ensure QoC requirements. The developer must explicitly program QoC requirements.

5.2 System Requirements

Digital phenotyping applications run in environments that have characteristics that can degrade QoC. Therefore, when designing these applications, it is necessary to consider some aspects to ensure their quality. The following describes some requirements that these continuous monitoring applications must have. We incorporated all the mentioned requirements in the application developed in this case study.

- **Identify active sensors:** The system must identify when a sensor is no longer active.
- **Inference from situations:** The system must provide information related to the worker's health status based on the sensors' vital signs information and activities performed, coming from sensors with a certain degree of imprecision.
- **Mobility support:** The system must guarantee data delivery even considering intermittent connections.
- **Resource-saving:** The system should try to minimize battery consumption and network interface usage of devices used by the worker when performing context data collection and distribution.

5.3 Specification of QoC Requirements

After identifying the requirements necessary to guarantee the quality of the application, the specification of the QoC requirements was carried out based on the parameters contemplated by the proposed metamodel. Figure 7 presents, in summary, how the application's QoC requirements were specified.

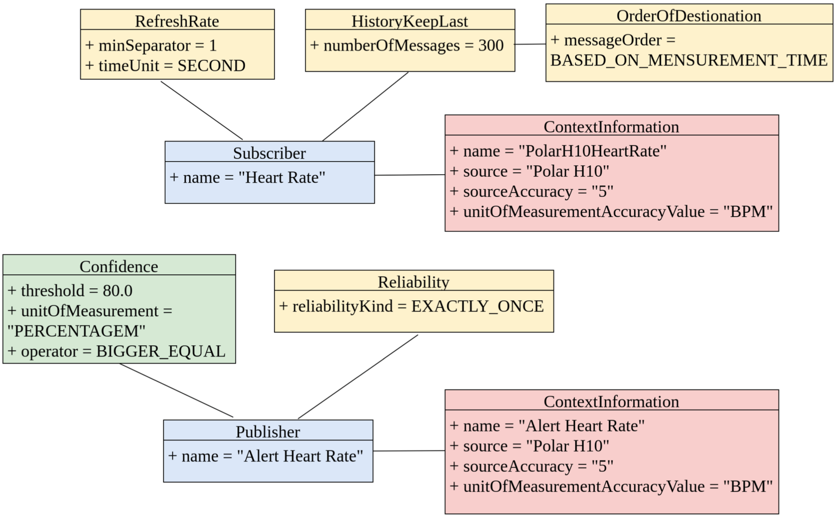


Fig. 7. Specification of QoC requirements for services that consume (Subscriber) heart rate information and disseminate (Publisher) heart rate alert information.

The monitoring application allows connection to devices that have Bluetooth technology. Bluetooth has a distance limit to keep an active connection. In the monitoring environment, in which the monitored user can move away from the smartphone that receives data from the device, it is necessary to notify consumer applications about the status of these services. The application developed in the case study publishes context information from several sensors: heart rate, breathing, blood pressure, and oxygen saturation. For each sensor, the Liveness parameter was defined. This parameter indicates that the service that publishes the sensor information will notify the consuming applications, informing them if they are still active or not.

Health professionals accompanying workers during working hours want to receive near real-time alerts when standard vital signs such as normal, altered, or critical ranges are detected. These alerts are essential to provide the professional with an indicator of the worker's current state of health for a quick reaction depending on the case's urgency. They are generated based on information from physical sensors. As the information from these sensors has a certain degree of imprecision, the information derived has a certain degree of confidence. We calculate the confidence of the alerts via an algorithm that measures the degree of confidence of a situation inferred based on the imprecision of the information source presented in [3]. In addition, the application provides an alert for heart rate (AlertHeartRate) and breathing rate (AlertBreathingRate), as these data are collected continuously. As specified in the requirements, to avoid false alerts about the actual situation of the worker's health status, the mobile application disseminates only the alerts with a degree of 80% confidence to the dashboard. It should be noted that continuous information must be generated by the sensor every 1 s, as specified by the RefreshRate parameter, for quick decision-making.

The mobile monitoring application uses the Activity Recognition API[4] developed by Google to identify the worker's activity. This API, in addition to providing the activity, also provides the confidence level of the situation. To ensure that the activity provided really matches the activity performed by the worker, a threshold of 80% was defined for the confidence parameter. As a result, the monitoring application will only receive situations with a confidence level of 80%.

Worker mobility can cause intermittent connection. This is a problem when you have essential items that must be delivered to a consumer application, such as alerts and information that is not collected continuously (e.g., blood pressure, glucose, and oxygen saturation). It is worth mentioning that a conventional sensor collects the blood glucose data, and the information is manually entered into the application. As specified in the requirements, the Reliability parameter is defined for this information to guarantee the delivery of the data. Each service that consumes context information has a History to store the information when there is no established internet connection or server connection. The Historic performs the function of a Buffer. Upon reestablishing the connection, the service disseminates the stored information.

Mobile devices have energy limitations due to the use of batteries that need to be recharged. The health monitoring system uses the user's smartphone to collect, infer and distribute information. In a standard data distribution model, the information is sent to the measure made available by the sensors. With each new data, network access is requested to send it. As per the Android documentation, each request to the network interface generates a reasonable power consumption. Therefore, for the worker monitoring application, a LatencyBudget parameter was defined that defines a delay for sending the data in a grouped way. As specified in the requirements, the delay is defined as 60 s, that is, the application will group the data generated in this interval and time and send them. This results in the number of network interface request the application makes.

6 Conclusion and Future Work

This study proposed a process to incorporate QoC requirements in digital phenotyping applications. The process has five steps, they are: specification of requirements, the transformation of the specification into target code, deployment of the code in the application, evaluation, and monitoring of the requirements, and finally, the visualization of the monitoring logs in a dashboard tool. From the proposed process, the study conceives a metamodel used to specify QoC requirements considering the structure of digital phenotyping applications in acquiring and distributing context information.

---

[4] https://developers.google.com/location-context/activity-recognition.

We have developed a system with QoC requirements for monitoring workers during working hours. It is worth mentioning that the specification transformation process was performed manually by the developers. The platform has two main applications. The first is a mobile application that collects signal and activity information from workers. The second is a dashboard that provides health professionals with near real-time information and historical data on the worker's health status. Through the case study, it was possible to observe that based on the proposed metamodel, it is possible to formalize QoC requirements that guarantee the quality of applications for digital phenotyping in the health field. In future works, we propose the automatic transformation of the specification into target code, evaluation and monitoring of the specified requirements, and dealing with conflicting situations between the specified QoC requirements.

Acknowledgments. This work was supported by the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES - Finance code 001); INCT of the Future Internet for Smart Cities (CNPq 465446/2014-0, CAPES 88887.136422/2017-00, and FAPESP 14/50937-1 and 15/ 24485-9); National Council for Scientific and Technological Development - CNPq (311608/2017-5, 420907/2016-5, 312324/2015-4); and the State of Maranhão Research Funding Agency - FAPEMA (via grant BM QUOTA UFMA - AGREEMENT N° 01/2021 FAPEMA/UFMA).

References

1. Aharony, N., Pan, W., Ip, C., Khayal, I., Pentland, A.: Social fMRI: investigating and shaping social mechanisms in the real world. Pervasive Mob. Comput. **7**(6), 643–659 (2011)
2. Bellavista, P., Corradi, A., Fanelli, M., Foschini, L.: A survey of context data distribution for mobile ubiquitous systems. ACM Comput. Surv. (CSUR) **44**(4), 1–45 (2012)
3. Bezerra, E.D.C., Teles, A.S., Coutinho, L.R., da Silva e Silva, F.J.: Dempster-shafer theory for modeling and treating uncertainty in IoT applications based on complex event processing. Sensors **21**(5), 1863 (2021)
4. Bolchini, C., et al.: And what can context do for data? Commun. ACM **52**(11), 136–140 (2009)
5. Buchholz, T., Küpper, A., Schiffers, M.: Quality of context: what it is and why we need it. In: Workshop of the HP OpenView University Association (2003)
6. Cho, S., Ensari, I., Weng, C., Kahn, M.G., Natarajan, K.: Factors affecting the quality of person-generated wearable device data and associated challenges: rapid systematic review. JMIR Mhealth Uhealth **9**(3), e20738 (2021)
7. Ferreira, D., Kostakos, V., Dey, A.K.: AWARE: mobile context instrumentation framework. Front. ICT **2**, 6 (2015)
8. Gomes, B.D.T.P., et al.: A middleware with comprehensive quality of context support for the internet of things applications. Sensors **17**(12), 2853 (2017)
9. Harel, D., Rumpe, B.: Modeling languages: Syntax, semantics and all that stu. N/A n/a, pp. 1–28 (2000)
10. Jagarlamudi, K.S., Zaslavsky, A., Loke, S.W., Hassani, A., Medvedev, A.: Requirements, limitations and recommendations for enabling end-to-end quality of context-awareness in IoT middleware. Sensors **22**(4), 1632 (2022)

11. Mendes, J.P., et al.: Sensing apps and public data sets for digital phenotyping of mental health: systematic review. J. Med. Internet Res. **24**(2), e28735 (2022)
12. Moura, I., et al.: Mental health ubiquitous monitoring supported by social situation awareness: a systematic review. J. Biomed. Inform. **107**, 103454 (2020)
13. Ometov, A., et al.: A survey on wearable technology: history, state-of-the-art and current challenges. Comput. Netw. **193**, 108074 (2021). https://doi.org/10.1016/j.comnet.2021.108074, https://www.sciencedirect.com/science/article/pii/S1389128621001651
14. Saccaro, L.F., Amatori, G., Cappelli, A., Mazziotti, R., Dell'Osso, L., Rutigliano, G.: Portable technologies for digital phenotyping of bipolar disorder: a systematic review. J. Affect. Disord. **295**, 323–338 (2021)
15. Schueller, S.M., Begale, M., Penedo, F.J., Mohr, D.C.: Purple: a modular system for developing and deploying behavioral intervention technologies. J. Med. Internet Res. **16**(7), e3376 (2014)
16. Silva, M., et al.: Neighborhood-aware mobile hub: an edge gateway with leader election mechanism for internet of mobile things. Mobile Netw. Appl. **27**(1), 276–289 (2020). https://doi.org/10.1007/s11036-020-01630-3
17. Teles, A.S., et al.: Enriching mental health mobile assessment and intervention with situation awareness. Sensors **17**(1), 127 (2017)
18. Statista: Number of smartphone subscriptions worldwide from 2016 to 2027 (2022). https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/. Accessed 13 Apr 2022
19. Torous, J., Kiang, M.V., Lorme, J., Onnela, J.P., et al.: New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment. Health **3**(2), e5165 (2016)
20. Xiong, H., Huang, Y., Barnes, L.E., Gerber, M.S.: Sensus: a cross-platform, general-purpose system for mobile crowdsensing in human-subject studies. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 415–426 (2016)

Patient Data Work with Consumer Self-tracking: Exploring Affective and Temporal Dimensions in Chronic Self-care

Tariq Osman Andersen[1](✉), Jonas Fritsch[2], and Stina Matthiesen[1]

[1] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
tariq@di.ku.dk
[2] Digital Design, IT University of Copenhagen, Copenhagen, Denmark

Abstract. Emerging studies are reporting on the implications of self-tracked data in patients' everyday life and how it influences self-care activities in chronic care. The increased uptake of consumer wearable activity trackers in healthcare contexts and the wider application of advanced analytics is changing the temporal scope from 'past-centric' to 'future-centric' personal informatics. At the same time, a stream of research is making clear that experiences of emotion are constitutive of patient data work suggesting that the micro practices of engaging with personal data has an important affective dimension. We conducted an exploratory interview study with five chronic heart patients with an implanted cardiac device to conceptualize the data work, which is involved in making sense of self-tracked data from a consumer wearable activity tracker (Fitbit Alta HR). In this paper, we contribute to understanding patient data work as seven forms of micro practices: Verifying, Questioning, Motivating, Reacting, Accepting, Distancing, and Sharing. We discuss how these practices relate to temporal and affective dimensions of engaging with self-tracked data in chronic care and point to future research.

Keywords: Self-tracking · Self-care · Wearable activity trackers · Personal informatics · Affective computing

1 Introduction

Self-care technologies in chronic care have traditionally been regarded as tools for supporting disease-related matters [1]. However, during the last two decades there has been a boom in consumer self-tracking devices, originally designed for sport, leisure, and wellness. These wearable and mobile devices are increasingly becoming part of patients' self-management practices and the line between the realms of medicalized self-care technologies and consumer self-tracking technologies is gradually being blurred [2, 3]. Along with this pervasive access to consumer health applications and wearable activity trackers, the work of *producing*, *gathering*, *interpretating* and *using* health data is no longer a job solely upheld by healthcare professionals. Instead, individuals with chronic conditions can engage in various forms of patient health data management practices, which has been broadly described as "patient data work" [4, 5].

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 666–680, 2023.
https://doi.org/10.1007/978-3-031-34586-9_44

Outside the medical contexts, engagements with personal information and self-tracking data have been conceptualised into different stages including activities of *preparation*, *collection*, *integration*, *reflection* and *action*. Studies under the term "personal informatics" have given shape to a long-standing discourse and multiple contributions to understanding how people engage with new types of digital devices that allow for collection of multiple forms of personal data [6]. Moreover, studies of using wearable activity trackers in clinical care contexts have been studied in health informatics [7], often with a focus on acceptability and effect in clinical encounters like using Fitbit in cancer or cardiac care or self-tracking among patients with co-morbidity [2, 8, 9]. In Human Computer Interaction (HCI), studies have examined how participants used self-tracking technologies as intertwined with self-care practices for disease monitoring and fitness tracking in chronic care [3, 10, 11].

More recently, two streams of research have appeared, which give rise to new questions: one stream examines emotion and affect during patients' data intensive practices [12–14]. For example, it is known that patients' experiences in diabetes self-care practices have an important affective dimension, and that patients and relatives are bound to the emotional struggle moving between control, freedom, and anxiety [13]. Another stream of research, which is concerned with future-oriented personal informatics in health contexts [15–17], has started to examine the opportunities and implications of artificial intelligence and advanced analytics in the space of self-tracking and self-care. For example, one study found that supporting "anticipation" and engaging in prospective and proactive approaches to tracking can provide new opportunities for design, which is somehow an extension to the withstanding focus on supporting reflection on historic data.

With this paper, we wish to explore two questions that cut across these newer streams of research and connect with earlier studies of self-tracking by seeking to conceptualise the micro practices that emerge when chronic heart patients are exposed to consumer wearable activity data with a Fitbit device.

- *What forms of patient data work do chronic heart patients engage in during self-care activities when using a consumer wearable activity tracker?*
- *What are the temporal and affective dimensions of patients' self-tracking and self-care activities?*

We conducted a qualitative interview study with five participants who all have chronic heart disease and an implanted cardiac defibrillator (ICD) which is remotely monitored at the Rigshospitalet, Copenhagen University Hospital, Denmark. The study is an extension of a previous study where 27 chronic heart patients were invited to use a Fitbit consumer device for 3–12 months and were interviewed about their experiences of self-tracking during self-care.

2 Background

2.1 Self-tracking with Consumer Wearable Technologies in Everyday Chronic Care

People living with chronic conditions engage in self-care activities to manage their disease as part of everyday life [1]. Some activities are an extension to the medicalized part of their treatment, such as the day-to-day management of prescribed medication and self-monitoring of symptoms as well as using medical grade equipment like blood glucose monitors or home telemonitoring equipment to manage the disease [1, 18]. Other self-care activities are related more to the mundane character of living with a chronic condition such as responding to the psychological and emotional impact of the disease and undertaking lifestyle changes as well as getting support from informal caregivers, searching for information, or communicating with other patients with similar diseases.

Prior research in health informatics has investigated the experiences of patients when engaging in self-tracking using consumer wearable activity trackers and smartphone applications in chronic care contexts [7, 9, 19–21]. Several studies have considered acceptability and adoption in medical care, for example research on the acceptability and attitudes towards integrating fitness tracking with a Fitbit device into clinical care of men with prostate cancer [9]. Other research examined the use of self-tracking among patients with multiple chronic illnesses and found that there are multiple purposes for tracking and that some patients consider it as work to track their own data [8]. Research also considered the implications of sharing "patient-generated" data in different types of patient-clinician interactions [20] suggesting that data produced and collected by patients, including activity and biomedical data from consumer wearables, becomes health data when used as part of the formal and the informal disease management. More recently, a study found that the effects of self-tracking with a consumer wearable device (Fitbit) in chronic care constituted ambivalent experiences i.e., both negative and positive experiences such as gaining new insights from data in one moment and data evoking doubts in another moment [2].

In HCI, studies of self-tracking have united around the term personal informatics and have focused on systems that support data intensive practices like collection and reflection upon self-tracked data [6, 22]. Early studies were mostly oriented towards the "quantified self" and domains of general health, fitness, and behaviour change i.e., non-medicalized contexts [22]. Today, HCI studies are increasingly exploring the role of self-tracked data in more medicalized contexts such as diabetes [10], irritable bowel syndrome [11], and multiple sclerosis [3] where findings include unpacking of how participants used self-tracking technologies as intertwined with self-care practices for disease monitoring and fitness tracking.

Self-reflection or just "reflection" on self-tracked data has been a prominent theme and it has been argued that designing for reflection is just as important as designing for "experience" through interaction [23]. For example, there can be different purposes of reflection such as learning, prompting action, and self-development and there are different levels of reflection such as descriptive (i.e., revisiting events), explanatory, and transformative (i.e., a fundamental change in understanding which might ultimately lead to a change in practice). The idea of designing for reflection on self-tracked data suggests an orientation towards the past and historic moments of data collected less than a future-oriented one. However, the advent of artificial intelligence and advanced analytics has pushed the perspective in HCI and personal informatics from a mostly historically oriented perspective in self-tracking towards a more future-oriented perspective.

Emerging studies have started to explore future-centric personal informatics whereby self-tracking data are turned into prognostics and personalized predictions. Lee and others [15] explore the differences between what they call ‘past-centric’ and ‘future-centric’ and found opportunities for stress management when supporting individuals’ anticipation and engaging in prospective and proactive approaches to tracking. Similarly, Rho et al. [17] conducted an experiment on supporting people with predictive information to lose weight and found that future-oriented ‘consequence information’ had a more positive impact than traditional ‘performance information’ when self-tracking for weight loss. Others have explored opportunities in chronic self-care and experimented with personalized predictions to generate nutrition-driven, and real-time forecasts of blood glucose levels to support decision-making among diabetes type 2 patients. Desai et al. [16] developed a smartphone app called GlucOracle and evaluated its feasibility for facilitating nutritional decision-making and found that technologically savvy individuals with well-managed blood-glucose experienced forecasts to be unsurprising and rarely prompting action while individuals with limited health technology experience and knowledge of diabetes self-management found predictions to be insightful and encourage concrete changes in diet and blood glucose management. These recent studies provide a different take on self-tracked data in healthcare contexts and foreground a temporal perspective on data work as well as emphasizing the need to explore the prognostic role which consumer wearable data may have for chronic patients.

2.2 Emotion and Affect in Design of Self-care Technologies and HCI

When looking across studies of patients’ engagement with self-tracking technologies there is a growing body of literature that has turned to investigate the role of affect and the emotional implications of patient data work [4, 5, 24, 25]. Research has investigated patients’ emotional experiences around self-monitoring, for example the ways in which blood glucose data collected by patients with diabetes and their caretakers is tightly bound to the emotional struggle moving between control, freedom, peace of mind and anxiety, and the burden of dealing with technology [13]. Similarly, it is found that data tracking for fertility self-monitoring promotes the achievement of certain positive goals but may accentuate negative emotions such as feeling burdened or abandoned [12]. Others have studied patient experiences in cardiac device telemonitoring and found that not having access to data or feedback from clinicians can create anxiety and the feeling of uncertainty as well as emotional and life-changing impact, which in turn creates doubt, guilt, and concern [26]. Positive and negative affect can therefore co-exist and become present as emotional ambivalence, which studies have found in healthcare contexts and among quantified self-enthusiasts [2, 25, 27, 28]. Conflicting or ambivalent experiences appear constitutive of self-tracking including affective responses like “doubt, guilt, fear, shame, dismay, disappointment, and hesitation as well as joy, relief, excitement, enthusiasm, and pride” [28].

Research into emotion in HCI research is, however, not new. In particular, the notion of affect has played a prominent role in HCI and design since Picard's pioneering work on Affective Computing [29]. Initially, it was argued that Affective Computing would be "computing that relates to, arises from, or deliberately influences emotion or other affective phenomena" (ibid.). Nonetheless, the main agenda of Affective Computing has been formulated as making computers recognize or express emotions [30, 31]. This has led to critique within HCI and interaction design where it has been argued that this definition of emotion as a kind of transferable "information" is reductionist and does not fully encompass the complexity of the human emotional experience [31]. Instead, it was argued that an "interactional" approach to affect that lets people reflect on their emotional richness was needed [32]. Within this interactional approach, Höök has suggested the term 'affective loop' pointing to the way that affect and emotion can emerge and be supported through interactions with technology and data involving both body and mind [33]. Lately, the concept of "affective health" has been coined by Sanches et al. [14], to encompass the stream of HCI studies related to affective disorders such as depression, anxiety, and bipolar health issues. In this paper, we follow this turn towards continued engagement with affective interactions and the role of emotions as interactive properties of technology design [34]. We further advance recent research that points to a holistic and encompassing engagement with how affective interactions on a micro-level can lead to relational changes on a macro-level [35]. Here, affect is conceptualised as constitutive for human experience, and not only in affective disorders. We build on this approach to explore affective dimensions of self-tracking technologies that comes to have an intimate role in e.g., chronic self-care. We also consider the temporal i.e., historic versus future-oriented engagement with self-tracked data to explore the implications of patient data work in a time where application of predictive analytics is emerging.

3 Study Design and Method

This study explores patient data work and the micro practices chronic heart patients engage in with various information sources including but not limited to their bodily sensations, communication with health professionals, and wearable activity data using a Fitbit Alta HR device. The study is an extension of a former study where 27 ICD patients were invited to use a Fitbit wearable activity tracker for 3–12 months and share their experiences through three semi-structured interviews [2]. Five patients were recruited from the original study and the selection was carried out using purposive sampling. Our criteria for inclusion were based on having a mix of participants who had different illness severities and diverse uses and experiences of the Fitbit device. The original study received formal ethical approval by the Capital Region of Denmark's Committee for Health Research Ethics (no. H-19029475) and patients provided informed consent and were carefully instructed about their participation in the extension of the project, which had no intervention component. Five semi-structured interviews were conducted using an interview guide with four themes, which revolved around the patients' day-to-day prognostic (i.e., future-oriented) work and their use of activity data, their emotional labor, and their informational needs. In addition, the study added speculative considerations of the usefulness of predictions based on artificial intelligence of severe heart arrhythmia in a smartphone app. Two interviews were carried out in-person and three interviews were carried out over Zoom due to Covid-19. The duration of the interviews was between 43 min and 131 min. All interviews were transcribed verbatim. Data analysis was carried out collaboratively using an inductive qualitative approach based on constructing grounded theory [36] supported by the qualitative data analysis software NVivo 12 (QSR International, Melbourne, Australia).

4 Findings: Patient Data Work

We identified seven forms of patient data work that describe the micro practices that participants engaged in by relating their lived, bodily experiences with consumer self-tracking data. We explored how these data work practices were characterised by having an emotional dimension as well as a temporal dimension ranging from the here and now (situational) to the reflective (historic) and prospective (future-oriented).

4.1 Verifying

The most prevalent patient data work practice with Fitbit was verifying bodily felt symptoms. Several of the participants reported how they began using heart rate data and activity data to confirm or check for the alignment with their heart related symptom experiences. All the participants explained how they had already developed a sensitivity towards particular bodily sensations for recognizing emergent severe heart arrhythmia. Previous experiences had taught them how certain symptoms were anticipatory of upcoming events and how these symptoms functioned as cues for taking action. P4 explained that when he experienced severe “chest cramps” he knew “automatically” that severe heart arrhythmia was underway, and it is due time for calling an ambulance and P1 told several stories of how he learned that “dizziness” and “near fainting” were clear signs of severe arrhythmias: *“I’m 110% sure of VT [severe heart arrhythmia] because when I get the VTs I feel badly uncomfortable. I almost faint, they are very clear”.*

Despite having developed a form of bodily awareness about sensations and symptoms, several of the participants began to use wearable data during everyday situations to check their heart condition and thereby verifying their symptom experiences. One participant explained that severe heart arrhythmia can feel a bit different from time to time and that he has begun to use heart rate data to become more certain: *“I can see that the curve is even for a normal high heart rate - it is the same all the way. But the curve is, definitely, not even when you have VT. I use the watch to be absolutely sure because it does not always feel the same” (P3).* In this way, Fitbit data became a tool for verifying experiences of symptoms and bodily sensations “here and now” and supported a form of patient self-diagnosing. By keeping a personal log and using his smartwatch to verify bodily sensations of severe heart arrhythmia, P3 expressed how he had become more aware of emergent arrhythmia by turning to the combination of data and symptoms: *“I am becoming more and more aware of it. I experience it every time I get confirmed when it shows that it is exactly as I have felt it [by combining Fitbit data and personal log data]”.*

Similarly, P1 described how he used Fitbit to verify his symptoms of atrial fibrillation (AF), another type of severe heart arrythmia: *"It is AF when the heart spins with a pulse from 110–140. This is the area where things start to get critical and then it is time. I've started to keep an eye on my pulse when I begin to sweat, and I can feel it on my breathing" (P1).* He gave an example of waking up in the middle of the night feeling uncomfortable, sweating, and having difficulties breathing, and explained how he turned to the Fitbit device to verify his symptom experiences: *"The times I have woken up with it at night, I could see from the intervals [in the Fitbit data] that I had been lying with a high pulse for a very long time."* For him, the combination of symptom experiences and Fitbit data enabled him to verify bodily sensations but moreover supported his decision on what action to take: *"When I had those episodes during the night, I could knock them down with an extra beta blocker [prescribed heart medicine], right, and if the pulse does not calm down then I have to call 112 or 1813 [emergency telephone number]."*

Like P4 and P1, P2 began to use Fitbit in relation to her heart condition. She developed a daily routine of checking in to see that everything is okay: *"(...) a couple of times a day I go in and check what my heart rhythm is, if it is higher"* (P2). Similarly, P2 also began to consult her heart rate data to confirm or disconfirm her symptom experiences: *"If I feel that my heart has been throbbing a little, I can go in and check to see, okay, is there anything behind this, or not."* She explained how she used the wearable data to calm her when in doubt about her heart condition while being outside and on the go: *"At the same time it [FitBit] can also calm you. During a winter holiday in Thy I was out walking. We were four and two of them were walking really fast. I felt I could walk fast but I could feel that I could not keep up with this, but in reality, I could see on my app that I could do it without problems. So, in this way you could say that it calms you down"* (P2).

While alike in comparing symptoms and data for verification purposes, these examples point to quite different temporalities of use. The real-time verification of bodily felt symptoms was supportive in the situation, but moreover worked as cues for potential upcoming events by knowing what action to take. This is different from the retrospective verification work of P3, but both forms of verification were afforded by the data-body loop. Besides appropriating the use of the wellness-fitness tracker towards a disease diagnostic device, we also found that some participants used verification for providing reassurance and emotional comfort, which underlines the importance of the affective dimension of patient data work with consumer wearable data.

4.2 Questioning

Another type of patient data work that emerged was using self-tracking data to explore or seek answers to disease related questions. One participant, P1, described how he began to use Fitbit for exploring associations between behaviour change and his heart. He was concerned about exploring the effects that smoking may have on his severe heart arrhythmias. By experiment, he found that his average heart rate clearly decreased when no longer smoking: *"From the day I stopped and ten days onwards my average heart rate dropped by one per day"* (P1). From this active questioning along with his use of self-tracking, he reasoned that it was the "tangibility" of his heart rate data and the visualization that enabled his discovery: *"This tells me that smoking, along with several* *other factors, can push me to a place where I can get some abnormal and irregular heart rhythms. If I had not had this [Fitbit], then I wouldn't have been able to get these answers. I could, perhaps, feel that I had gotten it a little better, but it is the tangible and visual, that makes it happen for me"* (P1).

Consumer self-tracking data can, in turn, generate prompts for new questions that lead to individual discoveries. Led by curiosity, some participants engaged in keeping track of their wearable activity data on a daily basis. One participant explained how she used Fitbit a couple of times a day to look for abnormal patterns. By actively questioning the relation between her behaviour, bodily sensations, and past experiences, she discovered that there may be a connection between increased heart rate and normal activity: *"I can see if the heart rate is in the red area, and I know that I have not been out for a walk then I think: "I need to be a little bit careful". But, if I know I have been out for a long brisk walk, then there is an explanation, and everything is fine, and I won't expect any events will come unannounced"* (P2).

Likewise, P1 discovered that just before arrhythmias occur, there is a traceable disconnect between not being active and data showing an abnormally high heart rate. This led to increased concerns and made P1 take action and contact the hospital: *"I think my heart rate was relatively high when I was in the normal range of activity, right? And that made me contact the hospital where I said: "we will have to get this under control: either there is something wrong with my medication, it does not work, or I would like to have a responsible doctor to take a closer look"* (P1).

In this way, participants engaged in micro practices of continuously exploring and reflecting upon self-tracking data to seek unanticipated discoveries or generate some form of answers to pertinent questions that may support their self-care practices. While the work of questioning typically has a retrospective orientation, patients may be affected emotionally in the situation of discovery and decide to take action. Moreover, the participants' discoveries could turn in to patient knowledge that the participants would use for purposes of verifying bodily felt symptoms.

4.3 Motivating

Motivating relates to the ways in which the wearable activity data became integrated with exercising, which the participants, like most other people, considered as good for their heart and overall health condition. For some of the participants, engagement with steps and heart rate data was linked to a desire to stay in control of their health condition and afforded positive affect. For P2 who had never experienced a shock from her ICD, the Fitbit data encouraged her to ensure that the activities she engaged in were good for her. She described how she was motivated to go from walking 10.000 to 15.000 steps and how FitBit had gradually motivated her to improve her exercise behavior: *"[I]n the beginning it was just the steps and then it was like the steps and number of days in a week where I get enough exercise – and then it became interesting with that heart rate or need for sleep[...]"* (P2).

Similarly, for P1, the visual cues of exercising more and improving behavior like eating healthier, decreasing stress or quitting smoking, motivated and provided reasons for changing his behavior: *"These visual things they make me happy it seems that it's actually working that you're doing this. To me it's reinforcing the situation you're in, well* *okay it's motivating to continue. Because when I feel better maybe I can do more. I can visually see the effects […]"* (P1). Just like verifying emergent symptoms with activity data became a routine activity, several participants explained how checking heart rate, number of steps, and stress levels was motivation for keeping a positive outlook and encouraged being more active: *"I look at it every day. I can get all the information I want so yes, I use it a lot. It gives me all the information I really want. In relation to my number of steps, my heart and stress"* (P5).

While some patients, like P4, did not find the data useful for motivating exercise due to inaccuracy of the data that *"don't make sense"*, several of the participants used Fitbit and the self-tracking data as a motivational device, which in general supported positive affective loops such as feeling motivated or seeing that efforts led to positive health outcomes such as a lower average heart rate and thus, lower risk of heart arrhythmia.

4.4 Reacting

Reacting was another form of patient data work, which was triggered when patients were prompted with information about heart arrhythmia episodes or other information signifying certain changes in their heart condition. For example, P2 explained that one time she received a phone call from the ICD remote monitoring clinic where they told her that several severe heart episodes were detected. P2 reacted by looking into her calendar to understand the circumstances and the possible reasons for having the heart arrhythmia: *"When the hospital called me and said you had an episode at that and that time then I went back to my calendar and found that it happened during a tough meeting at work, possibly when I was fired. So, there have been some pinpoints like that where I have had a mentally hard time. It's my belief that when I'm under hard psychic pressure it can trigger heart fibrillation"*. In this way, external prompts with clinical information spurs certain considerations that can involve reasoning between past events and possible cues for actions. For P2, being prompted with information about certain heart episodes, initiated reflections, and reasoning, which resulted in increased awareness about how psychic stress may induce problematic episodes.

Reacting became particularly present in conversations with participants about opportunities of forecasting the risk of future arrhythmic heart episodes using the wearable and implantable data. The participants engaged in speculating about what it would mean to be notified and technology-prompted about increased risk of upcoming arrhythmic episodes. P3 considered being notified in due time would enable him to react and take appropriate action: *"Getting notified a few hours in advance would mean that I can react and do something"* (P3). Similarly, P2 speculated how a prompt with short-term prediction of upcoming arrhythmia could enable her to be in secure surroundings: *"Well, I would not drive a car and be behind the wheels. If I was in the danger zone for the next 48 h, to get a shock, then I would make sure I was not alone but with someone who could help me" (P2).* P1 speculated that he would react by considering the forecast against his bodily feeling and take appropriate action, somehow reversely verifying the data induced forecast: *"I think I would hold such a risk forecast up against how I feel right now, to see if it might go in that direction" (P1).*

While some data work practices were mostly patient-initiated e.g. verifying and questioning, we found reacting to be a distinct type of activity, alluded by prompts with external informational cues and imposing upon the patient some form of actualization of the disease. Reacting to data-induced forecasts could – in best cases – support self-care by enabling the participants to take appropriate action, but the accompanied emotional labor could overshadow the opportunity and lead to a worsened situation.

4.5 Accepting

While wearable activity data and data from the participant's implanted cardiac device could afford a positive outlook, we also found that coping with negative data could be seen as an active process of accepting. For some participants, accepting was integral to using self-tracking data and could set negative as well as positive affective loops in motion: *"If there are days where I have not moved enough. Then I got a bad conscience because then I kind of got reminded of it"* (P3). In this way biometrics and activity data are not "just" data that can lead to a bad consciousness. Instead, it is more pertinent what is at stake with inactivity for some chronic patients: *"Yes, but it is important to keep going. It is whether you are a heart patient or not. I feel it's even more important that I keep in shape. Especially because I know my illnesses. Whichever way it goes, there is no standby, that is, at some point, I will have to be transplanted. Then, as far as possible, I want to be in good shape or whatever I can now" (P3).* As the quote demonstrates, self-tracking data carries along emotional labor and managing concerns in relation to the current health condition and the clinical prognosis. Activity data from consumer devices can, when incorporated into self-care, act as reminders of how well or how poorly you manage, not only your daily goals for exercise or lifestyle change, but moreover the prognostic outlook your chronic heart condition.

4.6 Distancing

Most of the participants had a positive attitude toward accepting their heart condition and welcoming "good" and "bad" news emerging from implant and wearable data. Yet, some of the participants also emphasized 'distancing' as a strategy to cope with their disease. P4 only experienced severe heart arrythmia a few times and for him it has been a meaningful strategy to think less of his heart disease in his everyday life and only consider it when severe heart arrhythmia emerges: *"Well, I did have a lot of concerns about it when I was diagnosed with the heart disease. But, I have the heart that I have, and the psyche that I have. It's just my life condition and it's not something I go and fill my head with"*. Similarly, for P5, who has struggled with longer periods of being in a depressive state and crying a lot, now prefers to distance himself from reflecting too much on his heart condition: *"So you start with asking questions all the time like "why did it happen" and "what can we do to make you feel better". I believe that the worrying makes most people more sick than it makes them healthy. We are sick already, there is no reason to make us sicker"*.

P4 and P5, also considered the critical and negative emotional effects that short term predictions with activity data could generate. P5 speculated about the consequences it would have had for a recent trip to IKEA where he experienced a cardiac arrest. He explained that he preferred his bodily sensing and his personal know-how over relying on data and technology to verify his symptoms: *"If you can sense it and if you can feel* *that now it's on its way to something. And if you can handle it – well then you don't need to know it before it comes" (P5).* He believes that arrhythmia risk predictions based on data of upcoming severe heart arrythmia could lead to increased worrying: *"Personally, I would probably be skeptical: is it necessary for me to know? For me, I would not care. I think it would give more concerns than it would give me joy to know" (P5).* Similarly, P4 speculated that his reaction to technology-based prompts of risk prognosis would lead to more worrying and ultimately trigger arrhythmic episodes: *"I think, if I found out that in 14 days "you can count on getting an event" then it would in itself lead to more worrying or the anxiety itself would provoke a heart attack".* As such, some participants prefer – in some situations – to distance themselves from data-initiated actualization of their disease. For them, it is the experience that negative affective loops emerge from engaging too much with data that can infer something about their developments in their disease state.

4.7 Sharing

The participants reported that sharing their disease-related concerns prompted by data from the implanted and wearable devices with partners was an important part of their data work and self-care practices. P2 explained that sharing concerns about arrhythmic episodes with her husband, for example, happened last time they called from the clinic: *"After the clinic called in December, I quickly told it to my husband and my mother. I was like - whoops, I need to take care of myself".* Similarly, for P5 it is critical for self-care to involve close relatives in coping with the heart disease whenever arrhythmic episodes emerge: *"It will always, always be a good idea to also involve your relatives in one way or another".* For P3, his wife supports him when severe episodes arise and becomes a close partner for detecting and coping with severe arrhythmic events: *"Just after the fainting on Saturday, it still sits in me. Because when I fainted and woke up again, I actually didn't know if I got a shock from the ICD. But then my wife was next to me and could tell me that I did not."*

Sharing concerns about data becomes particularly pertinent when there are opportunities for taking counter measures and managing emergent episodes with heart arrhythmia. When engaging the participants in discussing the use of self-tracking data for arrhythmia risk prediction, several of the participants speculated that they would share it with their partners, like P5 explained: *"It will make really good sense, because then you are two who can act".* P2 considered involving her husband, daughter, or others nearby to co-manage the critical situation: *"If I had known that I was in dangerous risk then I might have told my husband 'you'll have go with me' or one of the riding girls down at the stable 'could you please go with me and help out' or tell my daughter on the road 'try to listen here, it's not so good'".* Similarly, P1 speculated about contacting his girlfriend if he learned about increased risk: *"I would maybe contact my girlfriend if I was not near her, and then say: 'Well the forecast looks kind of bad, I just have to do something'".* These examples describe that patient data work is not only individual but is oftentimes connected with care activities where partners, relatives and informal caregivers take part. It suggests a move beyond the individual relation to the coupling between the bodily-lived and self-tracking and implant data to a wider social sphere, most commonly consisting of the person and the person's partner.

5 Discussion

While the patients in this study are all proactive and engaged users of self-tracking technologies, the ways in which these patients interpret and use their data vary. Some patients log and use their data to critically question and follow the medical treatment they receive. Other patients use their data exploratorily to constantly test the ways in which they can optimize or improve their health condition. However, one thing that the patients have in common is that they engage in data work practices of verifying, questioning, motivating, reacting, accepting, sharing, and distancing in relation to coping and living with a chronic illness.

Looking at the findings, we can identify several temporal dimensions regarding the patients' use of their self-tracked Fitbit data. In particular, we can distinguish between practices related to collecting and reflecting on data to 1) understand *past* developments of e.g. your heart rate when doing exercises or when trying to make sense of events in the past as seen during activities of questioning or by being prompted by past events as in the examples of reacting, 2) in cases where the patients explore connections between in the *present* moments as in many of the verification examples, e.g. when trying to figure out whether a severe heart arrythmia is coming within the next few minutes and 3) when concerned about *future* actions as e.g. in the motivational practices. Naturally, we see crossovers e.g., when looking at past data to better envision future actions, as in the case of P2 when verifying arrhythmias and considering the action to take. We also see that these temporalities are often strongly related to making sense and taking action, whether by building an understanding of past patterns, by navigating a difficult situation in the moment, or by developing future strategies of coping and improving quality of life and avoid severe arrhythmic episodes.

A strong current underlying many of the presented data work practices has to do with the affective and emotional aspects of living with a disease that can be potentially life threatening, and which has often been initiated by a traumatic event. Examples from the questioning practices also suggest a relation between positive and negative affect and the ability to take action in everyday life. Here, we see different ways in which some patients use self-tracking data from consumer wearables to explore pertinent questions to generate answers that support their self-care practices. Related to this, we can also see that the reacting practices carry negative affect since they are often initiated by the health professionals or prompted by self-tracked data, rather than the patients themselves and can hence come as a surprise or in the form of unwanted news. The distancing practices show how patients sometimes feel the need to "escape" their condition, or data that reinforce negative aspects of their condition – even though this practice is not always tenable in the long run. Accepting practices are in a different state; here, we see how the negative affect sometimes surrounding the physical condition and bodily data is always in process – and that it is also possible to turn negative data into positive affect if you find a way to act upon it. Importantly, we see that patients developed very different affective attachments to the interplay between bodily symptoms, their wellbeing, and data, which can lead to reifying affective loops; if you get motivated, knowing that you are continuously doing better will be a positive factor. However, if you are constantly experiencing the data as a reminder that you are not doing enough or that you are in fact losing your health, it can become a reinforcing negative spiral. This calls for a very personalized strategy for understanding the different parameters and ways of presenting data to best suit the patients' needs.

6 Conclusion

In this paper, we have explored patient data work among chronic heart patients with an ICD using a consumer wearable activity tracker (Fitbit Alta HR). We found seven micro practices: Verifying, Questioning, Motivating, Reacting, Accepting, Distancing, and Sharing. As we demonstrated in our findings, patients living with a chronic illness are already emotionally burdened and their health conditions entail that even mundane activities such as driving a car or going to IKEA is not worry-free. As part of coping with their health condition, the patients have developed different practices of relating their bodily cues with their self-tracked data. The ways in which the patients speculate about their reaction to and use of short-term predictions might also open a path for exploring how the use of predictive health technologies may support everyday planning, which may also be easier to speculate about in contradiction to emotional reactions toward being told by an app that you are in high risk of a severe arrhythmic heart event. However, our results also showed that the effect of such predictive health results might differ from patient to patient; it can potentially introduce feelings of reassurance for some patients, while for other patients they risk prompting (unnecessary) concern.

References

1. Nunes, F., Verdezoto, N., Fitzpatrick, G., Kyng, M., Grönvall, E., Storni, C.: Self-care technologies in HCI: trends, tensions, and opportunities. ACM Trans. Comput.-Hum. Interact. **22**(6), 1–45 (2015)
2. Andersen, T.O., Langstrup, H., Lomborg, S.: Experiences with wearable activity data during self-care by chronic heart patients: qualitative study. J. Med. Internet Res. (7), e15873 (2020)
3. Ayobi, A., Marshall, P., Cox, A.L., Chen, Y.: Quantifying the body and caring for the mind: self-tracking in multiple sclerosis. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, pp. 6889–690. ACM (2017)
4. Bossen, C., Pine, K.H., Cabitza, F., Ellingsen, G., Piras, E.M.: Data work in healthcare: an introduction. Health Inform. J. **25**(3), 465–474 (2019)
5. Torenholt, R., Saltbæk, L., Langstrup, H.: Patient data work: filtering and sensing patient-reported outcomes. Sociol. Health Illn. **42**(6), 1379–1393 (2020)
6. Epstein, D.A., et al.: Mapping and taking stock of the personal informatics literature. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, New York, NY, USA, vol. 4, no. 4, pp. 1–38. ACM (2020)
7. Shin, G., et al.: Wearable activity trackers, accuracy, adoption, acceptance and health impact: a systematic literature review. J. Biomed. Inform. **93**(1), 103153 (2019)
8. Ancker, J.S., Witteman, H.O., Hafeez, B., Provencher, T., Van de Graaf, M., Wei, E.: "You get reminded you're a sick person": personal data tracking and patients with multiple chronic conditions. J. Med. Internet Res. **17**(8), e4209 (2015)
9. Rosenberg, D., Kadokura, E.A., Bouldin, E.D., Miyawaki, C.E., Higano, C.S., Hartzler, A.L.: Acceptability of Fitbit for physical activity tracking within clinical care among men with prostate cancer. In: AMIA Annual Symposium Proceedings, Bethesda, MD, USA, vol. 2016, p. 1050. American Medical Informatics Association (2016)

10. Mamykina, L., Mynatt, E., Davidson, P., Greenblatt, D.: MAHI: investigation of social scaffolding for reflective thinking in diabetes management. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, pp. 477–486. ACM (2008)
11. Chung, C.F., et al.: Boundary negotiating artifacts in personal informatics: patient-provider collaboration with patient-generated data. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, New York, NY, USA, pp. 770–786. ACM (2016)
12. Figueiredo, M.C., Caldeira, C., Reynolds, T.L., Victory, S., Zheng, K., Chen, Y.: Self-tracking for fertility care: collaborative support for a highly personalized problem. In: Proceedings of the ACM on Human-Computer Interaction, New York, NY, USA, pp. 1–21. ACM (2017)
13. Kaziunas, E., Ackerman, M.S., Lindtner, S., Lee, J.M.: Caring through data: attending to the social and emotional experiences of health datafication. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, New York, NY, USA, pp. 2260–2272. ACM (2017)
14. Sanches, P., et al.: HCI and affective health: taking stock of a decade of studies and charting future research directions. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, pp. 1–17. ACM (2019)
15. Lee, K., et al.: Toward future-centric personal informatics: expecting stressful events and preparing personalized interventions in stress management. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–13. ACM, New York, NY, USA (2020)
16. Desai, P.M., Mitchell, E.G., Hwang, M.L., Levine, M.E., Albers, D.J., Mamykina, L.: Personal health oracle: explorations of personalized predictions in diabetes self-management. In: Cui, W., Zheng, J., Lewis, B., Vogel, D., Bi, X. (eds.) Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, pp. 1–13. ACM (2019)
17. Rho, S., et al.: FutureSelf: what happens when we forecast self-trackers' future health statuses? In: Marshall, J., Tennet, P. (eds.) Proceedings of the 2017 Conference on Designing Interactive Systems, New York, NY, USA, pp. 637–648. ACM (2017)
18. Piras, E.M., Miele, F.: Clinical self-tracking and monitoring technologies: negotiations in the ICT-mediated patient–provider relationship. Health Sociol. Rev. **26**(1), 38–53 (2017)
19. Mercer, K., Giangregorio, L., Schneider, E., Chilana, P., Li, M., Grindrod, K.: Acceptance of commercially available wearable activity trackers among adults aged over 50 and with chronic illness: a mixed-methods evaluation. JMIR Mhealth Uhealth **4**(1), e4225 (2016)
20. Zhu, H., Colgan, J., Reddy, M., Choe, E.K.: Sharing patient-generated data in clinical practices: an interview study. In: AMIA Annual Symposium Proceedings, Bethesda, MD, USA, vol. 2016, p. 1303. American Medical Informatics Association (2016)
21. Ancker, J.S., Witteman, H.O., Hafeez, B., Provencher, T., Van de Graaf, M., Wei, E.: The invisible work of personal health information management among people with multiple chronic conditions: qualitative interview study among patients and providers. J. Med. Internet Res. **17**(6), e4381 (2015)
22. Li, I., Dey, A., Forlizzi, J.: A stage-based model of personal informatics systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, pp. 557–566. ACM (2010)
23. Fleck, R., Fitzpatrick, G.: Reflecting on reflection: framing a design landscape. In: Viller, SA., Kraal, B. (eds.) Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction, New York, NY, USA, pp. 216–223. ACM (2010)
24. Piras, E.M.: Beyond self-tracking: exploring and unpacking four emerging labels of patient data work. Health Informatics J. **25**(3), 598–607 (2019)

25. Ruckenstein, M., Schüll, N.D.: The datafication of health. Annu. Rev. Anthropol. **46**(1), 261–278 (2017)
26. Andersen, T.O., Andersen, PR., Kornum, A.C., Larsen, T.M.: Understanding patient experience: a deployment study in cardiac remote monitoring. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, New York, NY, USA, pp. 221–230. ACM (2017)
27. Marent, B., Henwood, F., Darking, M.: Ambivalence in digital health: co-designing an mHealth platform for HIV care. Soc. Sci. Med. **215**(1), 133–141 (2018)
28. Salmela, T., Valtonen, A., Lupton, D.: The affective circle of harassment and enchantment: reflections on the ŌURA Ring as an intimate research device. Qual. Inq. **25**(3), 260–270 (2019)
29. Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997)
30. Aboulafia, A., Bannon, L.J.: Understanding affect in design: an outline conceptual framework. Theor. Issues Ergon. Sci. **5**(1), 4–15 (2004)
31. Sengers, P., et al.: The enigmatics of affect. In: Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, New York, NY, USA, pp. 87–98. ACM (2002)
32. Boehner, K., DePaula, R., Dourish, P., Sengers, P.: Affect: from information to interaction. In: Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility, New York, NY, USA, pp. 59–68. ACM (2005)
33. Höök, K.: Affective loop experiences – what are they? In: Oinas-Kukkonen, H., Hasle, P., Harjumaa, M., Segerståhl, K., Øhrstrøm, P. (eds.) Persuasive Technology. Lecture Notes in Computer Science, vol. 5033, pp. 1–12. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-68504-3_1
34. Lottridge, D., Chignell, M., Jovicic, A.: Affective interaction: understanding, evaluating, and designing for human emotion. Rev. Hum. Factors Ergon. **7**(1), 197–217 (2011)
35. Fritsch, J.: Affective interaction design at the end of the world. In: Proceedings of DRS 2018: Catalyst, pp. 896–908. Design Research Society (2018)
36. Charmaz, K.: Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. Introducing Qualitative Methods Series (2006)

Personalizing mHealth Persuasive Interventions for Physical Activity: The Impact of Personality on the Determinants of Physical Activity

Alaa Alslaity(✉), Najla Amutari, and Rita Orji

Dalhousie University, Halifax, NS, Canada
{Alaa.alslaity,Najla.Almutari,Rita.orji}@dal.ca

Abstract. Behaviour changes and persuasive mobile health (mHealth) technologies have shown success in motivating people to be more active and engage in physical activity. Research has demonstrated that persuasive interventions perform better if they are theory-driven and personalized. Thus, various research has addressed personalizing mHealth technologies based on different aspects. However, the literature lacks studies on the moderating effect of personality traits on the determinants of physical activity as identified by the Health Belief Model. To fill this gap, we conducted a large-scale study of 430 participants' physical activity behaviour, associated determinants, and individuals' personality traits. We developed a general model showing how the determinants impact physical activity and a personality-based model exploring the moderating effect of personality. Then, we explored the differences between the two models, as well as between distinct personalities within the personality-based model. Our findings show that people of distinct personalities respond differently to the behaviour change determinants. Based on the results, the paper provides recommendations for designing personalized persuasive mHealth interventions for promoting physical activity.

Keywords: Mobile Health · Physical Activity · Personality Traits · Health Belief Model (HBM) · Persuasive Technology

1 Introduction

Being physically inactive is a cause of many health diseases, including diabetes [26], obesity [29], and cardiovascular [47]. Besides, researchers have found that regular physical activities are strongly associated with a reduced risk for severe COVID-19 outcomes [41]. Many health authorities have recommended adults engage in at least 150 min/week of moderate to vigorous physical activity (MVPA) [36]. Many countries have promoted this recommendation based on solid evidence that regular physical behaviour results in a broad range of health benefits [41]. However, many people do not have the motivation to engage in regular physical activity. Thus, several persuasive interventions have been proposed to increase individuals' likelihood of being more physically active.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 681–698, 2023.
https://doi.org/10.1007/978-3-031-34586-9_45

Software systems designed to change individuals' behaviour are called Persuasive Technologies (PT) [10]. Persuasive technologies have been used extensively in several behaviour change domains, including physical activity. Furthermore, it has become a consensus in the literature that a one-size-fits-all approach is insufficient, and the persuasive interventions should be tailored to different users' groups. Thus, several studies have discussed this issue and proposed potential solutions to personalize PT based on various theories and models. Among these models is the Health Belief Model (HBM) [39], which is developed to explain why people may or may not take action to prevent diseases or activities that cause health issues. It states that the likelihood that an individual will engage in a health-related behaviour is influenced by six determinants: Perceived Susceptibility, Perceived Severity, Perceived Benefit, Perceived Barrier, Cue to Action, and Self-efficacy. The HBM is one of the most widely applied health behaviour theories [14, 25, 32, 33]. Over several years, the HBM has been used to predict factors that affect people's behaviours, such as eating habits and physical activity and inform behaviour change intervention design. Therefore, researchers have used this model to tailor persuasive interventions based on different factors, such as age, gender, and culture.

The literature has suggested that tailoring persuasive interventions based on users' personalities is an effective approach to enhance the performance of these interventions [17]. This suggestion has been demonstrated in the health domain as well as other fields, such as games [4, 34] and eCommerce [2]. Nonetheless, there is a dearth of research on whether personality moderates the impact of the HBM's determinants on people's behaviour. This is essential for designing theory-driven interventions that are tailored to be appropriate for each individual depending on their personality type. This paper aims to fill this gap by exploring the impact of personality traits on the HBM behaviour change determinants. It also shows how to tailor persuasive mobile interventions to various personalities. The research is guided by two overarching research questions: (1) How does the impact of the HBM's behaviours determinants of physical activity vary across Personality Traits? (2) How can mobile persuasive interventions for promoting physical activity be tailored to individuals who are high in distinct personalities?

To answer this research question, we conducted a large-scale study of 430 participants. Following the recommendation of a previous study [1], we employed the six determinants of the HBM model along with a seventh determinant (namely, the Social Influence), which has been shown to be a strong determinant of physical activity. To distinguish participants' personalities, we employed the Big Five-Factor model (FFM) [30]. The FFM categorizes people's personalities based on five broad factors (or traits): Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

Data were collected through a survey study that involved three parts: participants' demographics, perception of the HBM determinants, and personality test. Using our data, we developed two models: a general model showing how the determinants impact physical activity and a personality-based model exploring the moderating effect of personality traits. Then, we explored the differences between the two models, as well as between distinct personalities within the personality-based model. Our findings show that people who are high in distinct personalities respond differently to the behaviour change determinants. Our findings reveal that the relation between the determinants and physical activity behaviour varies based on an individual's personality. For example, people who are high in Openness are mostly influenced by Perceived Benefit and Self-Efficacy, while Conscientiousness can be motivated by Self-efficacy and Social Influence, but Perceived Barrier demotivates them. People high in Extraversion emerged as the most influenced personality, with five determinants being significantly related to them. In contrast, Agreeableness and Neuroticism emerged as the least influenced personalities with no significant positive relationship with any determinants. Based on our findings and extensive examination of the literature, we map the determinants to their corresponding persuasive strategies for operationalizing them and provide recommendations for designing persuasive interventions tailored to different personalities. To the best of our knowledge, this study is the first to examine the relationship between health behaviour change determinants (identified by the HBM) and personality traits (identified by the FFM) to develop guidelines for tailoring persuasive interventions to promote physical activity.

The contributions of this work can be summarized as follows: 1) It applies the HBM to conduct a comparative investigation of the determinant of physical activity. 2) It investigates the moderating effect of personality traits on the HBM determinants of physical activity. And 3) It maps the HBM determinants to the personality traits and provides recommendations for designing persuasive systems that are personalized based on the big five personality traits and informed by the HBM determinants.

2 Background

Several studies have demonstrated that designing persuasive interventions based on well-established theories enhances these interventions and makes them more successful [7, 31, 40]. These theories help understand health behaviours, which, in turn, help in tailoring persuasive interventions based on behaviour change determinants [35]. This section presents an overview of the main concepts used in this work, the Five-Factor Model of personality and the Health Belief Model. This is followed by a review of persuasive interventions for behaviour change, with a focus on works related to the HBM and physical activity.

Five-Factor Model (FFM) of Personality. Humans are different in their characteristics. The extensive research about human behaviours has led to introducing several theories about human personality. Among these theories, the Five-Factor Model (FFM) [44] is the most widely accepted personality theory. It highlights a set of five factors, known as the Big-five personalities. These personalities are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. These five personality traits cover a wide range of personalities, and they are known by their relative stability throughout individuals' lives. Each of the five factors is described by many characteristics [30].

Personality is hypothesized to affect habits and behaviours [38]. Researchers have studied the impact of personality on individuals' physical activity. For instance, Chan et al. [5] examined the interactions of leisure-time, physical activity and personality traits on wellbeing, and whether such interactions vary between older adults in Hong Kong (HK) and older adults in the United Kingdom (UK). Based on the analysis of data obtained from 349 participants, the authors concluded that "personality needs to be considered when promoting and providing physical activity for older adults, although more research is needed to further explore how this can work effectively". Gacek et al. [11] investigated the personality-based determinants of physical activity. The study targeted Polish and Spanish physical education students. 219 Polish and 280 Spanish students participated in the study. The International Physical Activity Questionnaire (IPAQ) [43] was used in the study. The results revealed difference between personality traits. For instance, the level of total, vigorous, and moderate physical activity increased along with the increase in extraversion, while a decrease occurred along with the increase in neuroticism.

Health Belief Model (HBM). One of the oldest and most widely employed models of health behaviour promotion [32]. It is designed to predict and explain health-related behaviours. Specifically, HBM explains why people may (or may not) engage in health-related behaviours. It postulates that the likelihood that an individual will participate in health-related behaviour is influenced by six constructs (or determinants) [32], as follows: 1) *Perceived susceptibility*: assessment of the perceived risk for developing a health condition of concern. 2) *Perceived Severity:* individual's assessment of the consequence of contracting the health condition of concern. 3) *Perceived benefit*: individual's perception of the good things that could happen from undertaking specific behaviours. 4) *Perceived barrier*: perception of the obstacles and cost of behaviour change. 5) *Cue to action*: exposure to factors that prompt action. 6) *Self-efficacy*: individuals' confidence in their competence to perform the new health behaviour.

The HBM model has shown to be successful in designing several persuasive interventions for health behaviour change, particularly physical activity [14]. For instance, Jalilian et al. [19] studied the factors related to regular physical activity among Iranian medical college students based on the HBM. A study by King et al. [25] examined the impact of HBM on physical activity for college students. Specifically, the study examines whether college students' perceived benefits, barriers, cues, and vigorous physical activity involvement differed significantly based on several factors, including gender, grade level, parental encouragement, and peer encouragement.

Another study by Hoseini et al. [18] investigated the effect of an education plan based on the health belief model on the physical activity of females at risk of hypertension. The study's findings showed a significant increase in the physical activity levels two months after the intervention, confirming the efficiency of the HBM on the physical activity of women at risk for hypertension. A large-scale study was conducted by Orji et al. [33] to understand how health behaviour relates to gamers type. The study investigates gamers' eating habits and their HBM determinants of healthy behaviour. Based on the results and the differences between the models, the study proposed two approaches (general and personalized approaches) for effective persuasive game design.

A more recent study by Almutari and Orji [1] investigates the determinants of physical activity in collectivist cultures using Saudi Arabia as a case study. In addition, the study investigates the moderating effect of age and gender on the impact of the determinants. The study found that Perceived Severity, Cue to Action and Social Influence are the strongest determinants of physical activity in Saudi adults.

The HBM determinants have also been operationalized in several behaviour change apps. For instance, Lue et al. [27] deployed Cue to Action in a break prompting system (called Time for Break). The app enables people to set their desired work duration and prompts them to stand up or move. The app implemented cue to action through periodic notifications adjustable via personalized settings to allow people to set up their preferred work and break duration.

Hatami et al. [16] studied the impact of HBM-based educational resources on nutritional behaviour for cancer prevention. The study found that Self-efficacy, Severity, and Benefits were perceived to have a higher impact. It concluded that education plans based on HBM and implemented through multimedia could change nutritional beliefs and behaviours to prevent colorectal cancer.

The HBM model has also been introduced to more recent studies that concern behaviours related to the recently emerged COVID-19 corona virus. For instance, Jose et al. [20] used the HBM determinants to investigate and understand people's perception and preparedness towards the pandemic. Another study by Mahindarathne [28] adopted the HBM to identify factors that affect prevention behaviour against COVID-19. The study revealed that Perceived Benefits and Self-efficacy had a significant positive impact, while Perceived Barriers had a significant negative impact. Accordingly, the study reinstates the usability of the HBM in exploring health behaviour.

These studies and others have shown the effectiveness of the HBM model on predicting factors that influence several health behaviours, including physical activity and informing intervention design. However, there is hardly any research investigating the moderating effect of personality on the impact of these determinants, especially in the area of physical activity. This paper aims to fill this gap by studying the moderating effect of the big five personalities on the impact of the HBM determinants on physical activity.

3 Study Design

Our study follows a quantitative research approach. To acquire the data needed for our study, we conducted a survey that assessed the impact of the determinants of HBM along with the Social Influence factor. This section discusses the study design regarding instruments used to evaluate the determinants and the personality test, participants, and data analysis.

3.1 Measurement Instrument

We conducted a large-scale survey to study the relationship between determinants that motivate physical activity and personality traits. The survey was developed after an extensive literature review of behaviour change theories, personality traits, physical activity behaviour motivators, and persuasive technology interventions for physical activity. The survey was also pilot tested on 15 participants for refinement. It is worth mentioning that this study was approved by the Research Ethics Board at the University. This survey instrument consists of three sections: participants' demographics, the behaviour change determinants, and personality assessment.

In the demographic section, we asked participants about age, gender, and education level. In the second section (behaviour change determinants), we relied on the six determinants of the HBM and the Social Influence. The HBM is used because the literature shows that HBM is a helpful framework for designing both long and short-term behaviour change interventions [13]. Also, it has been successfully adapted and deployed in many persuasive interventions for health [22, 45]. Based on the recommendations and the findings of previous studies [1], we also added Social Influence as a seventh determinant. This part of the survey involves seven sections: one for each determinant. A 7-point Likert scale ranging from "1 = Strongly disagree" to "7 = Strongly agree" was used for each question. All the survey questions were adapted from previous research where they were validated [9, 14, 21, 32, 35]. These HBM determinants questions include 1) eighteen questions measuring Perceived Benefits – e.g., being physically active most of the time would be beneficial to me; 2) fourteen questions measuring Perceived Barriers – e.g., A major barrier to physical activity for me is cost; 3) two questions measuring Perceived Susceptibility – e.g., If I do not stick to regular exercise, I will be at high risk for some physical inactivity related diseases; 4) three questions measuring Perceived Severity – e.g., The thought of ending up in the hospital due to physical inactivity related diseases scares me; 5) fourteen questions measuring Cue to Action – e.g., I am motivated to exercise if I gain weight and not fit in my clothing; 6) six questions measuring Self-efficacy – e.g., If I want, I could easily exercise within the next two weeks; and 7) four questions measuring Social Influence – e.g., I will be more physically active if my friend goes to the gym regularly. More sample questions are provided in the appendix.

In regards to the personality test, we used the 10-item personality inventory (BFI-10), a validated instrument for personality traits evaluation that has been widely employed [37]. The ten items are evaluated in a five-point Likert scale ranging from "1: Strongly disagree" to "5: Strongly agree".

3.2 Participants

Before starting the recruitment process, the research was approved by the ethics board at Dalhousie University. This ethics approval confirm the protection of participants privacy and information confidentiality. We recruited participants through universities email lists and posters published to the public on social media, such as Facebook and Twitter. All participants participated voluntarily, and no compensation was given to them. Information confidentiality we received a total of 442 responses, of which 12 were excluded due to incompleteness and wrong response to the attention determining questions. Among these participants, 264 are female, 164 are males, and two are not specified. Their ages range from 18 to 65.

3.3 Data Analysis

The data analysis was done using SmartPLS, a software for structural equation modelling (SEM) using the partial least squares (PLS) path modelling method [48]. SmartPLS is robust and efficient for analyzing complex relationships such as the one investigated in this paper. It has been used extensively by previous research and shown to be effective [1, 23, 32, 42].

In order to confirm that the collected data fits the model (i.e., whether the data replicates the seven determinants in physical activity behaviour), we conducted a component-based Confirmatory Factor Analysis (CFA) [8] using SmartPLS 3. Each indicator (question) loaded onto its corresponding factors. We retained only indicators that had factor loadings of at least 0.5 in the data [15]. In the next step, we used Partial Least Squares (PLS) Structural Equation Modeling (SEM) to establish the relationship between the seven determinants and the physical activity behaviour of different personalities. To do so, we developed a model showing the relationship between personality, the HBM determinants, and the likelihood of physical activity behaviour. Figure 1 shows the model structure.

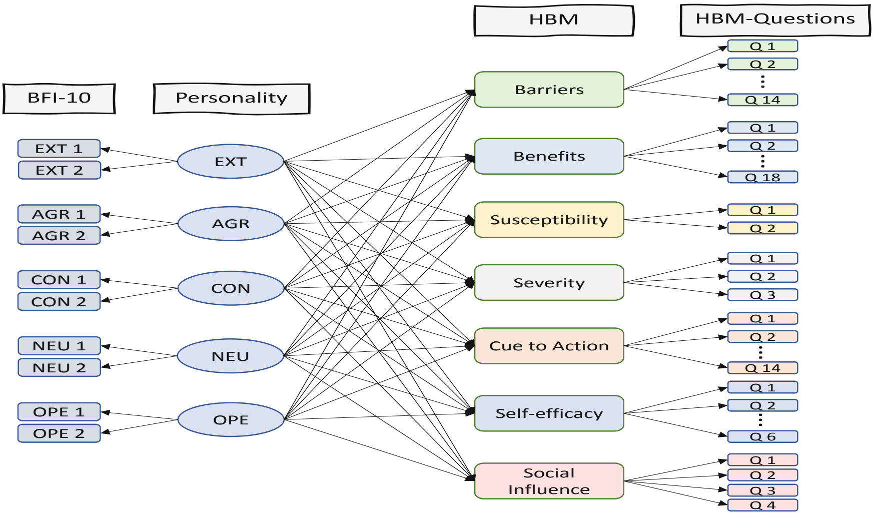


Fig. 1. PLS-SEM model structure

3.4 Measurement Validity and Reliability

Data reliability and validity were checked following the same approach implemented by previous research [1, 32, 34]. Specifically, data reliability was assessed using Cronbach's alpha and composite reliability scores. These measures show the strength of the correlation between indicators and their variables [32]. The results show that the indicators are reliable because Cronbach's alpha and composite reliability scores are higher than the threshold of 0.7 [6, 12]. Regarding data validity, it was checked using both convergent and discriminate validity. Data validity and reliability were satisfied for all the required criteria for the PLS-SEM. All constructs have an AVE (the Average Variance Extracted by the variables from its indicator items) above the recommended threshold of 0.5 [6]. The heterotrait-monotrait ratios of correlations (HTMT) were all below the recommended limit of 0.9.

4 Results

This section presents the results of our study, and it provides a discussion about our findings. The results are presented based on structural models, which determine the relationship between the determinants (susceptibility, Severity. Benefit, barrier, cue to action, self-efficacy, and Social Influence) and the behaviour. In particular, we have two structural models: one for the whole sample without considering the impact of personality (called the general model) and one showing the moderating effect of personality traits (we called it the personality-based model). In structural models, there are two important criteria; the level of the path coefficient (β) and the significance of the path coefficient (p) [15], where path coefficients measure the influence of a variable on another.

4.1 The General Model

This section discusses the relation between the HBM determinants and the likelihood of physical activity for the whole sample without considering the impact of personality. The results presented in Table 1 show that Cue to Action, Social Influence, Self-efficacy, and Perceived Severity emerged as significant motivators of physical activity. Specifically, Cue to Action emerged as the strongest determinant of physical activity behaviour overall. This is followed by Social Influence and Self-efficacy in the second and third place and Perceived Severity in the fourth place. On the other hand, Perceived Barrier emerged as the only determinant negatively associated with physical activity behaviour for the general model. This means that emphasizing this determinant in intervention design may demotivate people from engaging in physical activity behaviours. Finally, Perceived Benefit and Perceived susceptibility have no significant effect on people's physical activity behaviour.

Table 1. Standardized path coefficients and Significance of the general model (whole sample). The numbers represent significant coefficients at $p < .05$, and dash (-) represents non-significant coefficients.

BAR	BEN	CUA	EFF	SEV	SUS	SI
−0.19	-	0.25	0.21	0.11	-	0.22

BAR: perceived barrier, BEN: perceived benefit, CUA: cue to action, EFF: self-efficacy, SEV: perceived Severity, SUS: perceived susceptibility, SI: social influence

These results demonstrate the impact of each determinant on the whole sample. As mentioned above, it has been well established that the HBM determinants can predict factors influencing health behaviours without considering the impact of personality. The question that arises here, however, how generalizable are these results? For instance, can we conclude from Table 1 that Perceived barriers, benefits, and susceptibility should not be employed for promoting physical activity for all user types? The literature has shown that determinants' influence level can be moderated by several factors, including culture, age, gender, and gamer types [1, 32, 33]. We hypothesize that personality could moderate the extent to which the determinants influence physical activity behaviour. This hypothesis was motivated by the fact that personality has been shown to influence many aspects of people's lives, their beliefs, how to use technology, and their worldview [2, 3, 24, 34]. Nonetheless, there is hardly any research on the possible moderating effect of personality on the impact of the HBM determinants, especially in the area of physical activity. The next section shows how personality traits moderate the impact of the HBM determinants.

4.2 Moderating Effect of Personality Traits

This section discusses the relationship between personality traits and the HBM determinant. To achieve the research objective, we developed a model showing the relationship between personality, the HBM determinants, and the likelihood of physical activity behaviour, see Fig. 1. The individual path coefficients obtained from the model are summarized in Table 2. The numbers presented in the table show the level of path coefficients that are significant ($p < 0.05$), while the dashes (-) represent non-significant coefficients. Again, personality is a type scale, hence an individual cannot be classified as belonging to a single personality rather they could be high in one personality trait and low in others. Hence, we simultaneously modeled the relationship between personality traits and HBM factors as shown in Fig. 1. Accordingly, whenever we mention personality traits, we mean people who are high in the corresponding trait. For instance, mentioning "Openness people" indicates people who are high in openness facets according to the personality test (BFI-10).

Table 2 shows that the five personality traits are different with respect to how the HBM determinants impact their physical activity behaviours. People who are high in Openness are positively related to Perceived Benefits and Self-Efficacy, while Consciousness people are positively associated with Self-efficacy and Social Influence and negatively related to Perceived Barrier. Extraversion emerged as the most influenced personality; it is positively associated with five determinants (Perceived Benefit, Cue to Action, Perceived Severity, Perceived Susceptibility, and Social Influence). On the other hand, Agreeableness and Neuroticism emerged as the least susceptible personalities as the results did not show any significant positive association between them and the HBM determinants. However, Self-efficacy and Social Influence are negatively associated with Neuroticism.

By comparing the results of the general model (Table 1) with the personality-based model (Table 2), we notice how personality traits moderate the influence of the seven determinants. For instance, the general model shows that both Perceived Benefit and Perceived Susceptibility are non-significant motivators for the whole sample. However, the personality-based model shows that Perceived Benefits emerged as a significant motivator for Openness and Extraversion, while Perceived Susceptibility is a significant motivator for Extraversion. Also, the general model suggests that Social Influence and Self-efficacy are associated positively with physical activity behaviour. Although this is true for Conscientiousness and Extraversion, this is not the case for the other personalities. Particularly, Neuroticism is negatively associated with Social Influence and Self-efficacy. Besides, the general model shows that a total of four determinants (Cue

Table 2. Standardized path coefficients and significance for each Personality Trait. Dash (-) represents non-significant coefficients.

Factors	BAR	BEN	CUA	EFF	SEV	SUS	SI
Openness	-	0.14	-	0.12	-	-	-
Conscientiousness	(−0.35)	-	-	0.21	-	-	0.13
Extraversion	-	0.25	0.28	-	0.15	0.26	0.11
Agreeableness	-	-	-	-	-	-	-
Neuroticism	-	-	-	(−0.10)	-	-	(−0.10)

to Action, Self-efficacy, Perceived Severity, and Social Influence) are significantly associated with physical activity behaviour. However, the personality-based model revealed that the personalities perceive these determinants differently. For instance, Cue to Action (which emerged as the strongest determinant in the general model) is a strong determinant only for Extraversion, likewise Severity.

These results show that personality moderates the influence of the HBM determinants on physical activity behaviour. People with different personalities perceive the seven HBM determinants differently, highlighting the need to personalize the determinants in persuasive intervention to individuals' personality types.

5 Discussion and Design Recommendations

The results presented in the previous section demonstrate that personality traits moderate the impact of the HBM determinants on physical activity. In summary, Extraversion emerged as the most influenced personality (it has a significant relationship with five determinants). On the other hand, Agreeableness and Neuroticism emerged as the least influenced personalities. To a high extent, these conclusions are in line with previous studies, which have also demonstrated that Extraversion is the most responsive to persuasive strategies, and Neuroticism is the least responsive [34]. This section discusses our findings regarding the seven determinants. Based on these findings, Sect. 5.1 provides design recommendations for personalizing persuasive interventions for promoting physical activities.

Perceived Barriers. The general model shows that Perceived Barrier is negatively associated with physical activity. However, the personality-based model indicates that only Conscientiousness is negatively associated with Perceived Barriers. People high in Conscientiousness could be demotivated by any persuasive intervention that emphasizes perceived barriers associated with physical activities. A possible explanation of this negative relation is that individuals high in Conscientiousness are efficient and planful. Also, they are responsible and tend to be reliable and thorough. Since they are responsible and reliable, they tend to do things perfectly, and therefore they avoid any source of obstruction or hindrance that may lead to imperfect work. Thus, exposing people high in Conscientiousness to barriers can influence them negatively; this is in line with the

HBM proposition. The results also show that other personalities do not show a significant response to Perceived Barriers.

Cue to Action. Triggers of a target behaviour are important for promoting healthy behaviours. The results from our models demonstrate that Cue to Action is an important motivator of physical activity for people high in *Extraversion* only. This result was surprising given the diverse cues employed in persuasive interventions, such as reminders, prompts, and alerts [32]. Also, previous studies found that Cue to Action is an effective way for promoting healthy behaviour for different groups (e.g., gamers types [33] and collectivist and individualist groups [1]). Nonetheless, this is actually an interesting finding because it reveals the impact of personality traits on the effectiveness of persuasive intervention employing the determinants to promote physical activity. Our results show how different personalities can be influenced differently. Besides, other studies also found that Cue to Action did not promote health behaviour in interventions [31, 32].

Self-efficacy. Our results revealed that Self-efficacy is significantly positively associated with Openness and Conscientiousness personalities. As mentioned above, individuals high in Conscientiousness are efficient, planful, and reliable. Therefore, they tend to be confident in their abilities and efficiency as they follow plans. Regarding Openness, individuals who are high in Openness are usually defined as open to experience because they often seek new and unfamiliar experiences. Thus, they have a high level of self-confidence to explore new and unfamiliar things. On the other hand, Self-efficacy demotivates people high in Neuroticism. This negative impact can be explained by the fact that people high in Neuroticism are characterized by several negative and unstable moods. These moods lead Neurotic individuals to interpret typical situations as threatening (i.e., they may respond negatively to ordinary situations) [46]. Thus, as Table 1 shows, two determinants are perceived as negative by people high in Neuroticism, but none of the determinants significantly impact them positively.

Perceived Benefit. Our results show that Perceived Benefit is a significant positive determinant for *Openness* and *Extraversion* personalities. This finding can be explained by two points. First, people high in Openness tend to be intellectually curious, and they are more inclusive in their thinking than other persons. Therefore, they are more inclusive in assessing the benefits of their actions. Thus, they are significantly influenced by the perceived benefit determinant. Second, regarding people high in Extraversion, they are known as active, energetic, outgoing, and impulsive by nature. They prefer to *do* activities rather than *think* about doing the activities. Thus, they are significantly influenced not only by Perceived Benefit but also by Cue to Action, Perceived Severity, Perceived Susceptibility, and Social Influence.

Social Influence. Social Influence emerged as a strong positive determinant for *Conscientiousness* and *Extroversion.* This result was expected, especially because individuals high in Extraversion are talkative, friendly, and social. Thus, they are highly influenced by social-related factors. On the other hand, Social Influence emerged as a significant determinant that impacts *Neuroticism* negatively. As mentioned above, this negative association can be explained by the negative and unstable emotions that distinguish neuroticism personality.

Perceived Severity. We found that Perceived Severity is significantly associated with Extraversion only. This positive association implies that individuals people high in Extraversion care about the negative consequences of being physically inactive. As mentioned above, our results show that Extraversion is significantly associated with five determinants. Previous studies have also demonstrated that Extraversion emerged as the most responsive personality trait to persuasive strategies [34].

Perceived Susceptibility. Similar to Perceived Severity, the Perceived Susceptibility did not emerge as a strong determinant for any personality other than *Extraversion.* As mentioned before, because individuals high in Extraversion are active and enthusiastic, they are more likely to engage in new activities even without motivation. Their perception of the associated risk would increase the likelihood of their engagement in a new activity (physical behaviour in particular). On the other hand, Perceived Susceptibility is not associated negatively with any personality trait.

5.1 Design Recommendations

The previous section discusses how different determinants are associated with personality traits. This section builds on these results and provides recommendations for designing persuasive systems that are personalized based on the big five personality traits and informed by the HBM determinants. To do so, we relied on the established mapping between the HBM determinants and persuasive strategies introduced by Almutari and Orji [1] and depicted in Fig. 1. The mapping was done for the domain of promoting physical activity, and it was accomplished with the help of seven experts from several disciplines, including persuasive computing, human-computer interaction, and health.

Table 3. Sample mapping of determinants to persuasive strategies [1]

Determinant	Persuasive Strategies
Perceived Barriers	Suggestion, Extinction, Punishment, Negative reinforcement
Perceived Benefits	Reward, Gain-framed appeal
Cue to Action	Reminder, Suggestion
Self-efficacy	Incremental goal setting, Recognition, Feedback, Praise
Perceived Severity	Punishment, Negative reinforcement, Vicarious reinforcement, Simulation
Perceived Susceptibility	Self-monitoring, Loss-farmed appeal, Simulation
Social Influence	Cooperation, Social Facilitation, Social learning, Comparison

Based on these mappings, and the mapping between Personality traits and HBM (depicted in Fig. 1), we provide the following recommendations for designing personalized persuasive interventions for promoting physical activity (Table 4).

Table 4. Mapping determinants to personality traits. "√" indicates strategies that can be used, "X" indicates strategies that should be avoided, and "-" indicates no strong correlation was found.

	BAR	BEN	CUA	EFF	SEV	SUS	SI
Openness	-	√	-	√	-	-	-
Conscientious	X	-	-	√	-	-	√
Extraversion	-	√	√	-	√	√	√
Agreeableness	-	-	-	-	-	-	-
Neuroticism	-	-	-	X	-	-	X

Conscientiousness. Individuals high in *Conscientiousness* are motivated mainly by two determinants, Self-efficacy and Social Influence. Thus, to motivate *conscientious* people to be physically active, **we recommend that designers deploy persuasive strategies that promote users' confidence in their ability to perform healthy behaviours (i.e., self-efficacy-related strategies) or strategies that use the power of social influence**. A summary of these strategies is presented in Table 3. On the other hand, participants high in Conscientiousness are negatively influenced by Perceived Barriers. Therefore, designers **should avoid using strategies that allude to barriers associated with physical activity**. Strategies such as Negative reinforcement, punishment, and extinction should be avoided.

Extraversion. Our study shows that people high in Extraversion are the most influenced and easily motivated to engage in physical activity. Specifically, five determinants (Perceive Benefit, Cue to Action, Perceived Severity, Perceived Susceptibility, and Social Influence) emerged as strong motivators of people high in Extraversion. On the other hand, none of the determinants was found to have a significant negative impact on people high in Extraversion. These associations indicate that individuals high in Extraversion are motivated to be physically active by their perceptions of the good things related to being physically active (perceived benefit) and the risks and seriousness of the consequences of not being physically active (Severity). Therefore, **to design persuasive interventions targeted at promoting physical activity among people who are high in extraversion, designers could employ the benefit, cue to action, severity, susceptibility, and social influence related strategies.** Based on the established mapping (Table 3), using persuasive strategies, such as Rewards, Punishment, simulation, and self-monitoring, would increase the likelihood that extroverts will engage in physical activity. In addition, people high in Extraversion are motivated by factors that prompt physical activity (such as reminders and suggestions) and by social influence strategies (e.g., cooperation, social comparison, or social facilitation).

Neuroticism. Our findings demonstrate that none of the determinants significantly motivate people high in *Neuroticism* positively. On the other hand, two determinants (self-efficacy and social influence) negatively influence *Neuroticism* personality. Hence, any persuasive strategy that alludes to self-efficacy and social influence may negatively affect neurotic individuals. **Therefore, designers should avoid using persuasive**

strategies, such as Goal Setting, Feedback, Cooperation, or Comparison, in persuasive intervention that promotes physical activities for people high in Neuroticism. Again, designers should investigate other determinants that will motivate neurotics to be physically active.

Openness. People who are *Open* to experience (i.e., high in Openness) emerged as significantly positively influenced by two determinants, Perceived Benefit and Self-efficacy, and they are not influenced negatively by any determinant. **Therefore, to design persuasive interventions that promote physical activity among people who are high in Openness, designers could employ persuasive strategies associated with perceived benefit (e.g., Reward and Gain-Framed Appeal) and self-efficacy (e.g., Incremental Goal Setting, Feedback, Praise, and Recognition).**

Agreeableness. People high in Agreeableness are significantly associated with none of the seven determinants. That means none of the seven HBM determinants can significantly motivate people high in *Agreeableness* to be physically active. Hence, **we recommend that research explore more agreeableness-oriented determinants that can significantly motivate them to be physically active.** Another possible implication is that the HBM determinants do not generalize and cannot predict the likelihood of health behaviour for everyone.

6 Limitations

Despite our findings that can inform the design and development of persuasive applications for physical activity behaviour, there are limitations to applying our results. First, like most large-scale population-based research, our study relied on self-reported data that could be biased and may not accurately describe peoples' actual behaviour. That is, what is measured is peoples' belief rather than their actual behaviour. Second, although our work is based on a large-scale study, and the HBM model has been widely used in several health domains, we cannot confirm the validity of our models in domains other than physical activity behaviour. Therefore, applying our models' results in other domains should be done with caution. As part of our future work, we will apply our guidelines described above to design and evaluate a persuasive intervention tailored to the personality type.

7 Conclusions and Future Work

The literature has demonstrated the ability of the Health Belief Model (HBM) determinants to predict factors influencing health behaviours. It has also shown that personality influences many aspects of our lives. However, there is barely any research investigating whether personality moderates the impact of the HBM determinants on health behaviour, specifically in the area of physical activity. This work is a step toward filling this gap and developing a personalized persuasive mobile intervention. We conducted a large-scale study of 430 participants to explore the relationship between the big five personalities, the seven HBM determinants, and the likelihood of physical activity. Our model results revealed differences between personality traits and the HBM determinants. These differences indicate that personality traits moderate the impact of the HBM determinants on physical activity behaviour. Hence, there is a need to tailor the HBM determinants to an individual's personality in persuasive intervention design. *Extraversion* emerged as the most influenced personality, with five determinants being strongly related to it. On the other hand, *Agreeableness* and *Neuroticism* are the least influenced, with no significant positive relationship with any of the determinants. Based on our findings and extensive study of the literature, we provided recommendations for designing persuasive interventions tailored to different personalities to motivate them to be physically active. As part of our future work, we will apply our guidelines described above to design and evaluate a persuasive intervention tailored to the personality type. Besides, we will also study the effect of users' demographics on the determinants of physical activities.

References

1. Almutari, N., Orji, R.: Culture and health belief model: exploring the determinants of physical activity among Saudi adults and the moderating effects of age and gender. In: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, pp. 138–146 (2021). https://doi.org/10.1145/3450613.3456826
2. Alslaity, A., Tran, T.: The effect of personality traits on persuading recommender system users. In: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, IntRS 2020, pp. 48–56 (2020)
3. Alslaity, A., Tran, T.: Users' responsiveness to persuasive techniques in recommender systems. Front. Artif. Intell. **4** (2021). https://doi.org/10.3389/FRAI.2021.679459
4. Arteaga, S.M., González, V.M., Kurniawan, S., Benavides, R.A.: Mobile games and design requirements to increase teenagers' physical activity. Pervasive Mob. Comput. **8**(6), 900–908 (2012). https://doi.org/10.1016/J.PMCJ.2012.08.002
5. Chan, B.C.L., Luciano, M., Lee, B.: Interaction of physical activity and personality in the subjective wellbeing of older adults in Hong Kong and the United Kingdom. Behav. Sci. **8**, 8 (2018). https://doi.org/10.3390/BS8080071
6. Chin, W.: The partial least squares approach to structural equation modeling. Modern Methods Bus. Res. **295**(2), 295–336 (1998). Retrieved October 7, 2021 from https://books.google.ca/books?hl=en&lr=&id=EDZ5AgAAQBAJ&oi=fnd&pg=PA295&dq=The+partial+least+squares+approach+for+structural+equation+modeling.+Modern+methods+for+business+research&ots=49sD5nr-ip&sig=KmYZ-1y0JWAlfJF8drlKhZHR2ko#v=onepage&q=The partiallea
7. Consolvo, S., McDonald, D.W., Landay, J.A.: Theory-driven design strategies for technologies that support behavior change in everyday life. In: Proceedings of the Conference on Human Factors in Computing Systems, pp. 405–414 (2009).https://doi.org/10.1145/1518701.1518766
8. Costello, A., Osborne, J.: Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Practical Assess. Res. Eval. **10**, 1–7 (2019). https://doi.org/10.7275/jyj1-4868
9. Anderson, E.S., Wojcik, J.R., Winett, R.A., Williams, D.M.: Social-cognitive determinants of physical activity: the influence of social support, self-efficacy, outcome expectations, and self-regulation among participants in a church-based health promotion study. Health Psychol. Official J. Div. Health Psychol. Am. Psychol. Assoc. **25**(4), 510–520 (2006). https://doi.org/10.1037/0278-6133.25.4.510

10. Fogg, B.J.: Persuasive technology: using computers to change what we think and do. Persuasive Technol. Using Comput. Change What We Think Do 1–282 (2003).https://doi.org/10.1016/B978-1-55860-643-2.X5000-8
11. Gacek, M., Kosiba, G., Wojtowicz, A., López Sánchez, G.F., Szalewski, J.: Personality-related determinants of physical activity among Polish and Spanish physical education students. Front. Psychol. **12**, 6360 (2022). https://doi.org/10.3389/FPSYG.2021.792195
12. Gefen, D., Straub, D., Boudreau, M.-C.: Structural equation modeling and regression: guidelines for research practice. Commun. Assoc. Inf. Syst. **4**(1), 7 (2000). https://doi.org/10.17705/1CAIS.00407
13. Glanz, K.: Theory at a glance: a guide for health promotion practice (1997). Retrieved October 5, 2021 from https://books.google.ca/books?hl=en&lr=&id=rUXiaSxFT48C&oi=fnd&pg=PA5&dq=Theory+at+a+glance:+A+guide+for+health+promotion+practice&ots=xlqpC3qDpc&sig=RaqlCqM7BUrHLUX30syAkJuege4
14. Gristwood, J.: Applying the Health Belief Model to Physical Activity Engagement Among Older Adults. Illuminare 9 (2011). Retrieved October 7, 2021 from https://scholarworks.iu.edu/journals/index.php/illuminare/article/view/1035
15. Hair, J.F., Ringle, C.M., Sarstedt, M.: PLS-SEM: indeed a silver bullet. J. Mark. Theory Practice **19**(2), 139–152 (2014). https://doi.org/10.2753/MTP1069-6679190202
16. Hatami, T., Noroozi, A., Tahmasebi, R., Rahba, A.: Effect of multimedia education on nutritional behaviour for colorectal cancer prevention: an application of health belief model. Malaysian J. Med. Sci. MJMS **25**(6), 110 (2018). https://doi.org/10.21315/MJMS2018.25.6.11
17. Hirsh, J.B., Kang, S.K., Bodenhausen, G.V.: Personalized persuasion: tailoring persuasive appeals to recipients' personality traits. Psychol. Sci. **23**(6), 578–581 (2012). https://doi.org/10.1177/0956797611436349
18. Hoseini, H., Maleki, F., Moeini, M., Sharifirad, G.R.: Investigating the effect of an education plan based on the health belief model on the physical activity of women who are at risk for hypertension. Iran. J. Nurs. Midwifery Res. **19**(6), 647 (2014). Retrieved October 10, 2021 from /pmc/articles/PMC4280731/
19. Jalilian, F., et al.: Predicting factors related to regular physical activity among Iranian Medical College student: an application of health belief model. Soc. Sci. 3688–3691 (2016). Retrieved October 10, 2021 from https://medwelljournals.com/abstract/?doi=sscience.2016.3688.3691
20. Jose, R., Narendran, M., Bindu, A., Beevi, N., Manju, L., Benny, P.V.: Public perception and preparedness for the pandemic COVID 19: a health belief model approach. Clin. Epidemiol. Global Health **9**, 41–46 (2021). https://doi.org/10.1016/J.CEGH.2020.06.009
21. Kasser, S.L., Kosma, M.: Health beliefs and physical activity behavior in adults with multiple sclerosis. Disabil. Health J. **5**(4), 261–268 (2012). https://doi.org/10.1016/J.DHJO.2012.07.001
22. Kharrazi, H., Faiola, A., Defazio, J.: Healthcare game design: behavioral modeling of serious gaming design for children with chronic diseases. In: Jacko, J.A. (ed.) HCI 2009. LNCS, vol. 5613, pp. 335–344. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02583-9_37
23. Khoi, B.H., Van Tuan, N.: Using SmartPLS 3.0 to analyse internet service quality in Vietnam. In: Anh, L.H., Dong, L.S., Kreinovich, V., Thach, N.N. (eds.) ECONVN 2018. SCI, vol. 760, pp. 430–439. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73150-6_34
24. Komatsu, H., Nishio, K.I.: An experimental study on motivational change for electricity conservation by normative messages. Appl. Energy **158**, 35–43 (2015). https://doi.org/10.1016/j.apenergy.2015.08.029
25. Merianos, A.L.: Vigorous physical activity among college students: using the health belief model to assess involvement and social support. Arch. Exerc. Health Dis. **4**(2), 267–279 (2014). https://doi.org/10.5628/aehd.v4i2.153

26. Lemes, Í., et al.: Sedentary behaviour is associated with diabetes mellitus in adults: findings of a cross-sectional analysis from the Brazilian National Health System. J. Public Health (Oxf.) **41**(4), 742–749 (2019). https://doi.org/10.1093/PUBMED/FDY169
27. Luo, Y., Lee, B., Yvettewohn, D., Rebar, A.L., Conroy, D.E., Choe, E.K.: Time for break: understanding information workers' sedentary behavior through a break prompting system. In: Proceedings of the Conference on Human Factors in Computing Systems, April 2018 (2018). https://doi.org/10.1145/3173574.3173701
28. Mahindarathne, P.P.: Assessing COVID-19 preventive behaviours using the health belief model: a Sri Lankan study. J. Taibah Univ. Med. Sci. **16**(6), 914–919 (2021). https://doi.org/10.1016/J.JTUMED.2021.07.006
29. Martínez-Ramos, E., et al.: Patterns of sedentary behavior in overweight and moderately obese users of the Catalan primary-health care system. PLoS ONE **13**, 1 (2018). https://doi.org/10.1371/JOURNAL.PONE.0190750
30. McCrae, R.R., John, O.P.: An introduction to the five-factor model and its applications. J. Pers. **60**(2), 175–215 (1992). https://doi.org/10.1111/J.1467-6494.1992.TB00970.X
31. Michie, S., Johnston, M., Francis, J., Hardeman, W., Eccles, M.: From theory to intervention: mapping theoretically derived behavioural determinants to behaviour change techniques. Appl. Psychol. **57**(4), 660–680 (2008). https://doi.org/10.1111/J.1464-0597.2008.00341.X
32. Orji, R., Mandryk, R.L.: Developing culturally relevant design guidelines for encouraging healthy eating behavior. Int. J. Hum. Comput. Stud. **72**(2), 207–223 (2014). https://doi.org/10.1016/J.IJHCS.2013.08.012
33. Orji, R., Mandryk, R.L., Vassileva, J., Gerling, K.M.: Tailoring persuasive health games to gamer type. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2013).https://doi.org/10.1145/2470654
34. Orji, R., Nacke, L.E., Di Marco, C.: Towards personality-driven persuasive health games and gamified systems. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (2017). https://doi.org/10.1145/3025453
35. Orji, R., Vassileva, J., Mandryk, R.: Towards an effective health interventions design: an extension of the health belief model. Online J. Public Health Inform. **4**, 3 (2012). https://doi.org/10.5210/OJPHI.V4I3.4321
36. Piercy, K.L., et al.: The physical activity guidelines for Americans. JAMA **320**(19), 2020–2028 (2018). https://doi.org/10.1001/JAMA.2018.14854
37. Rammstedt, B., John, O.P.: Measuring personality in one minute or less: a 10-item short version of the big five inventory in English and German. J. Res. Pers. **41**(1), 203–212 (2007). https://doi.org/10.1016/J.JRP.2006.02.001
38. Rhodes, R.E., Rfaeffli, L.A.: Personality and physical activity. In: Acevedo, E. (ed.) The Oxford Handbook of Exercise Psychology, pp. 195–223. Oxford University Press (2012). Retrieved May 30, 2022 from https://books.google.ca/books?hl=en&lr=&id=VR1pAgAAQBAJ&oi=fnd&pg=PA195&dq=personality+and+physical+activity&ots=5Au_6mL0QO&sig=oDlzHOEqc3vhJudLoqIlztewaJ4#v=onepage&q=personalityandphysicalactivity&f=false
39. Rosenstock, I.M.: Why people use health services. Milbank Q. **83**, 4 (2005). https://doi.org/10.1111/J.1468-0009.2005.00425.X
40. Michie, S., Prestwich, A.: Are interventions theory-based? Development of a theory coding scheme. Health Psychol. Official J. Div. Health Psychol. Am. Psychol. Assoc. **29**(1), 1–8 (2010). https://doi.org/10.1037/A0016939
41. Sallis, R., et al.: Physical inactivity is associated with a higher risk for severe COVID-19 outcomes: a study in 48 440 adult patients. Br. J. Sports Med. **55**(19), 1099–1105 (2021). https://doi.org/10.1136/BJSPORTS-2021-104080

42. Sarstedt, M., Cheah, J.-H.: Partial least squares structural equation modeling using SmartPLS: a software review. J. Mark. Anal. **7**(3), 196–202 (2019). https://doi.org/10.1057/S41270-019-00058-3
43. Sebastião, E., et al.: The international physical activity questionnaire-long form overestimates self-reported physical activity of Brazilian adults. Public Health **126**(11), 967–975 (2012). https://doi.org/10.1016/J.PUHE.2012.07.004
44. Tupes, E.C., Christal, R.E.: Recurrent personality factors based on trait ratings. J. Pers. **60**(2), 225–251 (1992). https://doi.org/10.1111/J.1467-6494.1992.TB00973.X
45. Peng, W.: Design and evaluation of a computer game to promote a healthy diet for young adults. Health Commun. **24**(2), 115–127 (2009). https://doi.org/10.1080/10410230802676490
46. Widiger, T.A., Oltmanns, J.R.: Neuroticism is a fundamental domain of personality with enormous public health implications. World Psychiatry **16**(2), 144 (2017). https://doi.org/10.1002/WPS.20411
47. Young, D.R., et al.: Sedentary behavior and cardiovascular morbidity and mortality: a science advisory from the American Heart Association. Circulation **134**(13), e262–e279 (2016). https://doi.org/10.1161/CIR.0000000000000440
48. Product | SmartPLS. Retrieved October 7, 2021 from https://www.smartpls.com/

Author Index

A
Abdalazim, Nouran 18
Aflatoony, Leila 528
Aguirre, Maia 397
Ahmed, Ashir 496
Ahmed, Md. Sabbir 33, 460, 557
Ahmed, Nova 33, 460, 557
Alchieri, Leonardo 18
Alecci, Lidia 18
Alonso, Itziar 149
Alpano, Paul Vincent 274
Alslaity, Alaa 477, 681
Amir, Shajnush 33
Amutari, Najla 681
Andersen, Tariq Osman 666
Atiba, Samuelson 33

B
Babichenko, Dmitriy 603
Baglione, Anna N. 570
Balasingham, Ilangko 75
Baoluo, Cheng 161
Bardzell, Jeffrey 584
Bardzell, Shaowen 584
Barnes, Laura E. 570
Bergsland, Jacob 75
Bestek, Mate 361
Biehl, Jacob 603
Bosschaerts, Kyana 636
Bould, Helen 379
Bouton-Bessac, Emma 247
Brabenec, Lubos 259
Braun, Melinda C. 540
Burgess, Romana 379

C
Cabrera-Umpierrez, María Fernanda 149
Carrasco-Lopez, Carmen 93
Casier, Koen 344
Cavalieri, Stefano 149
Chan, Gerry 477
Chen, Liming 3
Christou, Maria 450
Cobo, Antonio 127
Consoli, Antonio 138
Conte, Raffaele 104
Coutinho, Luciano Reis 651
Crombez, Pieter 344
Culpin, Iryna 379

D
da Silva e Silva, Francisco José 651
Dapaah, Chloe 570
De Brouwer, Mathias 344
De Marez, Lieven 636
de Moura, Ivan Rodrigues 651
De Turck, Filip 344
Dias, Nervo Verdezoto 33
Distler, Meg 207
Drouin, Michelle 207

F
Farrahi, Katayoun 174
Fernández-Canosa, Sara 138
Fico, Giuseppe 149
Filippodou, Elizabeth 149
Flanagan, Mindy 207
Flierl, Markus 412
Fotiadis, Dimitrios I. 49, 450
Fritsch, Jonas 666

G
Garcia, Meritxell 397
Gashi, Shkurta 18
Gatica-Perez, Daniel 247
Gentzler, Ryan D. 570
Georga, Eleni I. 49, 450
González, Verónica López 138
Grönvall, Erik 361
Guha, Shion 429
Guo, Bin 3

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
A. Tsanas and A. Triantafyllidis (Eds.): PH 2022, LNICST 488, pp. 699–701, 2023.
https://doi.org/10.1007/978-3-031-34586-9

H
Hall, Richard D. 570
Hasan, Tanvir 557
Hashem, Tanzima 289
Hayes, Patti 207
Herrera, Jorelle Aaron 274
Ho, Chih-Chun 3

I
Imamura, Yosuke 496
Ingram, Katherine H. 60
Islam, Khandker Aftarul 289

J
Ji, Changqing 161
Jimeno, Daniel 93
Jing, Qin 161
Johnson, Danielle 570

K
Katehakis, Dimitrios G. 332
Katsarou, Daphne N. 450
Kazi, Tahsin 60
Keppitiyagama, Chamath 412
Kerrigan, Connie 429
Khanshan, Alireza 315
Knoche, Hendrik 619
Kondylakis, Haridimos 332
Konstantakopoulos, Fotios S. 49
Kostalova, Milena 259
Kouroubali, Angelina 332
Kovac, Daniel 259

L
La Rosa, Davide 104
Larraza, Joseba Aitzol Arbilla 18
Laurindo, Luís Eduardo Costa 651
Licitra, Lisa 149
López-Fernández, Jacobo 397
Lopez-Perez, Laura 149

M
Madeira, Rui Neves 513
Mahima, K. T. Y. 412
Mahootiha, Maryamalsadat 75
Mallavarapu, Aditi 603
Manos, Anastassios 149
Markopoulos, Panos 315
Martinelli, Elena 149
Matthiesen, Stina 666
McMahan, Deborah 207
Medina, Juan E. 127
Meegahapola, Lakmal 247
Mekyska, Jiri 259
Méndez, Ariane 397
Mercalli, Franco 149
Mottola, Luca 412
Muro, Almira Astrid 274

N
Nabney, Ian 379
Nadarajah, Stephanie Githa 619
Nelis, Jelle 344
Niranjan, Mahesan 174
Nóbrega, Rui 513

O
Olsson, Carl Magnus 93, 116, 439
Ongenae, Femke 344, 636
Orji, Rita 477, 681

P
Palumbo, Filippo 104
Papaloukas, Costas 450
Pater, Jessica A. 207, 429
Payyapilly Thiruvenkatanathan, Kamala 584
Pearson, Rebecca 379
Pedersen, Peder Walz 619
Pedrasa, Jhoanna Rhodette 274
Pérez-Borrego, Yolanda 93
Pérez-Rodríguez, Rodrigo 127
Peters, Nelson 207
Pfafman, Rachel 429
Pinnock, Hilary 116
Ponakaladinne, Kiran 60
Pozo, Arantza del 397

Q
Qadir, Hemin Ali 75
Qi, Jun 229

R
Rahman, Md. Mahfuzur 557
Rektorova, Irena 259
Robles-Mateos, Paula 127
Rodríguez-Laso, Ángel 127
Rodríguez-Mañas, Leocadio 127
Ronca, Alessandra 104

Rony, Rahat Jahangir 33
Roxo, Gonçalo 513
Rusu, Andrei C. 174
Ruyssinck, Joeri 344

S
Saha, Priyeta 289
Salvi, Dario 93, 116, 439
Sansone, Francesco 104
Santini, Silvia 18
Shah, Syed Ahmar 116
Shahriar, Hossain 60
Sixian, Luo 496
Skovfoged, Milo M. 619
Song, Wenchao 3
Soto-León, Vanesa 93, 138
Sparkes, Valerie 33
Stawarz, Katarzyna 33
Steenwinckel, Bram 344, 636
Stojchevska, Marija 344
Stragier, Jeroen 636
Sun, Guanhua 528
Sun, Zhuo 3

T
Tesconi, Mario 104
Tigas, Stelios 450
Tolman, Sarah 570
Tonacci, Alessandro 104
Toscos, Tammy 207, 429
Tsang, Kevin C. H. 116
Tuason III, Philip Luis 274

U
Unzueta, Luis 397

V
Valero, Maria 60
Van Der Donckt, Jeroen 344
van der Herten, Joachim 344
Van Gorp, Pieter 315
Van Hoecke, Sofie 344, 636
Van Ooteghem, Jan 344
Villalba-Mora, Elena 127
Voigt, Thiemo 412

W
Walker, Erin 603
Wang, Xiang 229
Wang, Zhu 3
Weerasekara, Malith 412
Wells, Kristen J. 570
Wen, Nannan 603
Wilson, Andrew M. 116
Wilson, Richard 477
Wilson, Wendell 528
Wölfel, Matthias 540

Y
Yang, Po 229
Yang, Yu 229
Ymeri, Gent 93, 439
Yu, Zhiwen 3
Yue, Peng 229

Z
Zhang, Hualei 3
Zhao, Liang 60
Zhiyong, Gao 161
Ziadeh, Hamzah 619
Zoysa, Kasun De 412
Zulfiker, Syed Muhammad Ibne 289
Zumin, Wang 161