

Ruidan Su

Yudong Zhang

Han Liu

Alejandro F Frangi *Editors*

# Medical Imaging and Computer-Aided Diagnosis

Proceedings of 2022 International  
Conference on Medical Imaging  
and Computer-Aided Diagnosis  
(MICAD 2022)



# Lecture Notes in Electrical Engineering

## Volume 810

### Series Editors

Leopoldo Angrisani, Department of Electrical and Information Technologies Engineering, University of Napoli Federico II, Naples, Italy

Marco Arteaga, Departament de Control y Robótica, Universidad Nacional Autónoma de México, Coyoacán, Mexico

Bijaya Ketan Panigrahi, Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, Delhi, India

Samarjit Chakraborty, Fakultät für Elektrotechnik und Informationstechnik, TU München, Munich, Germany

Jiming Chen, Zhejiang University, Hangzhou, Zhejiang, China

Shanben Chen, Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

Tan Kay Chen, Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore

Rüdiger Dillmann, Humanoids and Intelligent Systems Laboratory, Karlsruhe Institute for Technology, Karlsruhe, Germany

Haibin Duan, Beijing University of Aeronautics and Astronautics, Beijing, China

Gianluigi Ferrari, Università di Parma, Parma, Italy

Manuel Ferre, Centre for Automation and Robotics CAR (UPM-CSIC), Universidad Politécnica de Madrid, Madrid, Spain

Sandra Hirche, Department of Electrical Engineering and Information Science, Technische Universität München, Munich, Germany

Faryar Jabbari, Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA

Limin Jia, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Alaa Khamis, German University in Egypt El Tagamoa El Khames, New Cairo City, Egypt

Torsten Kroeger, Stanford University, Stanford, CA, USA

Yong Li, Hunan University, Changsha, Hunan, China

Qilian Liang, Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX, USA

Ferran Martín, Departament d'Enginyeria Electrònica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

Tan Cher Ming, College of Engineering, Nanyang Technological University, Singapore, Singapore

Wolfgang Minker, Institute of Information Technology, University of Ulm, Ulm, Germany

Pradeep Misra, Department of Electrical Engineering, Wright State University, Dayton, OH, USA

Sebastian Möller, Quality and Usability Laboratory, TU Berlin, Berlin, Germany

Subhas Mukhopadhyay, School of Engineering & Advanced Technology, Massey University, Palmerston North, Manawatu-Wanganui, New Zealand

Cun-Zheng Ning, Electrical Engineering, Arizona State University, Tempe, AZ, USA

Toyoaki Nishida, Graduate School of Informatics, Kyoto University, Kyoto, Japan

Federica Pascucci, Dipartimento di Ingegneria, Università degli Studi "Roma Tre", Rome, Italy

Yong Qin, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

Gan Woon Seng, School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore, Singapore

Joachim Speidel, Institut of Telecommunications, Universität Stuttgart, Stuttgart, Germany

Germano Veiga, Campus da FEUP, INESC Porto, Porto, Portugal

Haitao Wu, Academy of Opto-electronics, Chinese Academy of Sciences, Beijing, China

Walter Zamboni, DIEM - Università degli studi di Salerno, Fisciano, Salerno, Italy

Junjie James Zhang, Charlotte, NC, USA

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering—quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact [leontina.dicecco@springer.com](mailto:leontina.dicecco@springer.com).

To submit a proposal or request further information, please contact the Publishing Editor in your country:

#### **China**

Jasmine Dou, Editor ([jasmine.dou@springer.com](mailto:jasmine.dou@springer.com))

#### **India, Japan, Rest of Asia**

Swati Meherishi, Editorial Director ([Swati.Meherishi@springer.com](mailto:Swati.Meherishi@springer.com))

#### **Southeast Asia, Australia, New Zealand**

Ramesh Nath Premnath, Editor ([ramesh.premnath@springernature.com](mailto:ramesh.premnath@springernature.com))

#### **USA, Canada**

Michael Luby, Senior Editor ([michael.luby@springer.com](mailto:michael.luby@springer.com))

#### **All other Countries**

Leontina Di Cecco, Senior Editor ([leontina.dicecco@springer.com](mailto:leontina.dicecco@springer.com))

**\*\* This series is indexed by EI Compendex and Scopus databases. \*\***

Ruidan Su · Yudong Zhang · Han Liu ·  
Alejandro F Frangi  
Editors

# Medical Imaging and Computer-Aided Diagnosis

Proceedings of 2022 International Conference  
on Medical Imaging and Computer-Aided  
Diagnosis (MICAD 2022)


 Springer

*Editors*

Ruidan Su  
Department of Computer Sciences  
and Engineering  
Shanghai Jiao Tong University  
Shanghai, China

Han Liu  
College of Computer Science and  
Software Engineering  
Shenzhen University  
Shenzhen, Guangdong, China

Yudong Zhang   
Department of Informatics  
University of Leicester  
Leicester, UK

Alejandro F Frangi   
Computational Medicine  
University of Manchester  
Manchester, UK

ISSN 1876-1100

ISSN 1876-1119 (electronic)

Lecture Notes in Electrical Engineering

ISBN 978-981-16-6774-9

ISBN 978-981-16-6775-6 (eBook)

<https://doi.org/10.1007/978-981-16-6775-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

This work is subject to copyright. All rights are solely and exclusively licensed 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 Singapore Pte Ltd. The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Organization

## Honorary Chairs

Prof. James Duncan, Yale University, USA

Prof. Leo Joskowicz, The Hebrew University of Jerusalem, Israel

## Conference Chairs

Asst. Prof. Ruidan Su, Shanghai Jiao Tong University (SJTU), China

Prof. Yu-dong Zhang, University of Leicester, UK

## General Co-Chairs

Prof. Alejandro F Frangi, University of Manchester, UK

Prof. Moi Hoon Yap, Manchester Metropolitan University, UK

## Program Committee Chairs

Prof. Ryuji Hamamoto, Tokyo Medical and Dental University, Japan

Prof. Baiying Lei, Shenzhen University, Shenzhen, China

Asst. Prof. Han Liu, Shenzhen University, China

Prof. Dr. Joseph M. Reinhardt, The University of Iowa, IA, USA

## **Program Co-Chairs**

Asst. Prof. Xiaoxiao Li, The University of British Columbia, Canada  
Asst. Prof. Islem Rekik, Istanbul Technical University, Turkey

## **Technical Chairs**

Prof. Qingfeng Chen, Guangxi University, China  
Asst. Prof. Thi Hoang Ngan Le, University of Arkansas, USA  
Asst. Prof. Duygu Sarikaya, Gazi University, Turkey

## **Publicity Chairs**

Asst. Prof. Genovefa Kefalidou, University of Leicester, United Kingdom  
Dr. Shuihua Wang, University of Leicester, United Kingdom

## **Local Chair**

William Wang, University of Leicester, United Kingdom

## **Local Co-Chair**

Hunter Zhu, University of Leicester, United Kingdom

## **Endorsed By**

MICCAI Society

## **Technical Program Committees**

Prof. Zakaria Belhachmi, Université Haute-Alsace, France  
Prof. Qiang (Shawn) Cheng, University of Kentucky, USA

Asst. Prof. Hao Chen, The Hong Kong University of Science and Technology, Hong Kong SAR, China  
Dr. Sarada Prasad Dakua, Hamad General Hospital, Qatar  
Prof. Smain FEMMAM, IEEE senior member, University of Haute-Alsace France, France  
Prof. Dagan Feng, University of Sydney, Australia  
Dr. Linlin Gao, Ningbo University, China  
Prof. Juan Manuel Górriz, University of Granada, Spain  
Prof. Yuzhu Guo, Beihang University, China  
Asoc. Prof. Stathis Hadjidemetriou, Cyprus International Institute of Management  
Asst. Prof. Yuankai Huo, Vanderbilt University, USA  
Assoc. Prof. Jianming Liang, Arizona State University, USA  
Assoc. Prof. Abbas Khosravi, Deakin University, Australia  
Asst. Prof. Jérôme Lapuyade-Lahorgue, University of Rouen, LITIS, France  
Asst. Prof. Mingxia Liu, University of North Carolina at Chapel Hill, USA  
Asst. Prof. Cheng Lu, Case Western Reserve University, USA  
Ms. Jiahong Ouyang, Stanford University, USA  
Prof. Jianquan Ouyang Xiangtan University, China  
Prof. Xiang Pan, Jiangnan University, China  
Prof. Gemma Piella, Pompeu Fabra University, Spain  
Prof. Su RUAN, LITIS laboratory, University of Rouen, France  
Asst. Prof. MEHDI SALIMI, St. Francis Xavier University, Canada  
Prof. Ute Schmid, University of Bamberg, Germany  
Prof. Thomas Schultz, Institute of Computer Science II, University of Bonn, Germany  
Dr. Rachel Sparks, King's College London, United Kingdom  
Asst. Prof. Yanmei Tie, Harvard Medical School, USA  
Dr. Tatiana Tyukina, University of Leicester, United Kingdom  
Assoc. Prof. Jichuan Xiong, Nanjing University of Science and Technology, China  
Dr. Guang Yang, National Heart & Lung Institute, Imperial College London, United Kingdom  
Dr. Qingyu Zhao, Stanford University, USA  
Dr. Jun Zhuang, Indiana University-Purdue University at Indianapolis (IUPUI), USA



# Preface

Welcome to the proceedings of the International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD 2022), held at the prestigious University of Leicester, UK, from November 20th to 21st, 2022. We are delighted to present this proceedings that showcase the latest advancements in the fields of medical imaging and computer-aided diagnosis.

MICAD has long been recognized as an important conference series dedicated to fostering innovation and collaboration among researchers and practitioners in the realm of medical imaging and computer-aided diagnosis. With each passing year, the conference continues to grow in scope and significance, and MICAD2022 was no exception.

MICAD2022 received submissions from 33 countries, in total, 98 full papers, and each paper was reviewed by at least three reviewers in a standard single blind peer review process. After meticulous evaluation and deliberation, 47 outstanding papers were accepted for presentation at MICAD 2022. (acceptance rate of 48%).

The papers featured in this volume encompass a wide array of topics within the fields of medical imaging and computer-aided diagnosis. They represent the collective efforts of researchers from diverse backgrounds, united by their shared commitment to advancing the frontiers of knowledge in healthcare technology. We are confident that the insights and innovations presented in these papers will contribute significantly to the ongoing progress in these vital domains.

We extend our deepest gratitude to all the authors who submitted their work to MICAD2022, as well as to our dedicated reviewers for their rigorous assessments. Additionally, we would like to acknowledge the support and contributions of the organizing committee, and the program committee for hosting this event. Our hope is that the discussions, insights, and findings presented in these proceedings will inspire future research, collaborations, and innovations in the realm of medical imaging and computer-aided diagnosis.

Once again, we extend a warm welcome to you as you embark on a journey into the rich tapestry of cutting-edge research showcased within the proceedings.

Shanghai, China

Dr. Ruidan Su

# Contents

## Medical Imaging

<b>Optimizing the Non-local Means Filtering of CT Images</b> .....	3
Ivo Draganov and Veska Gancheva	
<b>Exploring Structure-Wise Uncertainty for 3D Medical Image Segmentation</b> .....	15
Anton Vasiliuk, Daria Frolova, Mikhail Belyaev, and Boris Shirokikh	
<b>Towards Developing a Lightweight Neural Network for Liver CT Segmentation</b> .....	27
Mohammed Yusuf Ansari, Snigdha Mohanty, Serah Jessy Mathew, Subhashree Mishra, Sudhansu Sekhar Singh, Julien Abinahed, Abdulla Al-Ansari, and Sarada Prasad Dakua	
<b>NuRISC: Nuclei Radial Instance Segmentation and Classification</b> .....	37
Esha Sadia Nasir and Muhammad Moazam Fraz	
<b>A Semi-supervised Framework for Automatic Pixel-Wise Breast Cancer Grading of Histological Images</b> .....	53
Kenglung Chang, Yanyuet Man, and Hailong Yao	
<b>Lunatum Prosthetic Replacement: Modeling Based on Volume Rendering of CT Scan Images</b> .....	67
Manal Hamda, Btihal El Ghali, Imane Hilal, Omar El Midaoui, Nabil Ngote, Bahia El Abdi, and Kawtar Megdiche	
<b>Augmented Reality Applications for Image-Guided Robotic Interventions Using Deep Learning Algorithms</b> .....	77
Jenna Seetohul, Mahmood Shafiee, and Konstantinos Sirlantzis	
<b>Transfer Learning Based Classification of Diabetic Retinopathy on the Kaggle EyePACS Dataset</b> .....	89
Maria Tariq, Vasile Palade, and YingLiang Ma	

**Ex-vivo Evaluation of Newly Formed Bone After Lumbar Interbody Fusion Surgery Using X-ray Micro Computed Tomography** ..... 101  
 Jakub Laznovsky, Adam Brinek, Tomas Zikmund, and Jozef Kaiser

**Community Detection in Medical Image Datasets: Using Wavelets and Spectral Methods** ..... 111  
 Roozbeh Yousefzadeh

**Non-pooling Network for Medical Image Segmentation** ..... 121  
 Weihu Song, Heng Yu, and Jianhua Wu

**Lung CT Analysis Using 3D Disparity-Regularised Block Matching for Stereotactic Ablative Body Radiotherapy** ..... 131  
 Durai Arun Pannir Selvam, David I. Laurenson, William H. Nailon, and Duncan B. McLaren

**Identification of Melanoma Diseases from Multispectral Dermatological Images Using a Novel BSS Approach** ..... 143  
 Mustapha Zokay and Hicham Saylani

**2.5D Lightweight Network Integrating Multi-scale Semantic Features for Liver Tumor Segmentation** ..... 155  
 Yilin You, Zhengyao Bai, Yihan Zhang, and Jiajin Du

**Registration of Medical Image Sequences Using Auto-differentiation** ..... 169  
 Tomas Vicar, Roman Jakubicek, Jiri Chmelik, and Radim Kolar

**Small Animal Imaging: Iterative Algorithms Combined with Regularization Schemes, an Application to a Dual-Head Small Animal PET** ..... 179  
 Evangelia Karali

**Early Detection of Parkinson’s Disease Dementia Using Dual-Sided Multi-scale Convolutional Neural Networks (DSMS-CNN)** ..... 191  
 Callum Altham, Huaizhong Zhang, Marcello Trovati, Ella Pereira, Nicola Ray, Simon Keller, Antonella Macerollo, and Hulya Wiesmann

**A Change Detection with Machine Learning Approach for Medical Image Analysis** ..... 203  
 Mauro Mazzei

**U-Net###: A Powerful Novel Architecture for Medical Image Segmentation** ..... 231  
 Firat Korkmaz

**Computer-Aided Detection/Diagnosis**

**Optimising Chest X-Rays for Image Analysis by Identifying and Removing Confounding Factors** ..... 245

Shahab Aslani, Watjana Lilaonitkul, Vaishnavi Gnananathan, Divya Raj, Bojidar Rangelov, Alexandra L. Young, Yipeng Hu, Paul Taylor, Daniel C. Alexander, NCCID Collaborative, and Joseph Jacob

**3D-3D Rigid Registration: A Comparative Analysis Study on Femoral Bone Scans** ..... 255

Perrine Solt, Adlane Habed, Antoine Bautin, Pierre Maillat, and Michel de Mathelin

**Fully Automatic Axial Vertebral Rotation Measurement of Children with Scoliosis Using Convolutional Neural Networks** ..... 269

Jason Wong, Marek Reformat, and Edmond Lou

**Diagnostic Accuracy and Reliability of Deep Learning-Based Human Papillomavirus Status Prediction in Oropharyngeal Cancer** .... 281

Agustina La Greca Saint-Esteben, Chiara Marchiori, Marta Bogowicz, Javier Barranco-García, Zahra Khodabakhshi, Ender Konukoglu, Oliver Riesterer, Panagiotis Balermipas, Martin Hüllner, A. Cristiano I. Malossi, Matthias Guckenberger, Janita E. van Timmeren, and Stephanie Tanadini-Lang

**Optimizing the Illumination of a Surgical Site in New Autonomous Module-based Surgical Lighting Systems** ..... 293

Andre Mühlenbrock, René Weller, and Gabriel Zachmann

**An Eye-Tracking Based Machine Learning Model Towards the Prediction of Visual Expertise for Electrocardiogram Interpretation** ..... 305

Mohammed Tahri Sqalli, Dena Al-Thani, Mohamed B. Elshazly, Mohammed Al-Hijji, Alaa Alahmadi, and Yahya Sqalli Houssaini

**Synthetic Data as a Tool to Combat Racial Bias in Medical AI: Utilizing Generative Models for Optimizing Early Detection of Melanoma in Fitzpatrick Skin Types IV–VI** ..... 317

Daniel Kvak, Eva Březinová, Marek Biroš, and Robert Hrubý

**BD-Transformer: A Transformer-Based Approach for Bipolar Disorder Classification Using Audio** ..... 331

Mohamed Ramadan, Hazem Abdelkawy, Mustaqueem, and Alice Othmani

**Establishment and Analysis of a Combined Diagnostic Model of Acute Myocardial Infarction Based on Random Forests and Artificial Neural Networks** ..... 343  
Zhenrun Zhan, Xiaodan Bi, Jinpeng Yang, Xu Tang, and Tingting Zhao

**Striped-Cross Attention Network with Implicit Semantic Knowledge for Antibody Structure Prediction** ..... 353  
Miao Gu and Min Liu

**A Mobile Monitoring Application for Post-traumatic Stress Disorder** ..... 365  
Sirine Chaari, Chaima El Ouni, and Alice Othmani

**COVID-19 Diagnosis and Classification from CXR Images Using Vision Transformer** ..... 377  
Md Mahbubur Rahman, Shihabur Rahman Samrat, Abdullah Al Ahad, Mahmud Elahi Akhter, Ibraheem Muhammad Moosa, Rajesh Palit, and Ashfia Binte Habib

**Improved Techniques for the Conditional Generative Augmentation of Clinical Audio Data** ..... 389  
Mane Margaryan, Matthias Seibold, Indu Joshi, Mazda Farshad, Philipp Frnstahl, and Nassir Navab

**Learning from Failure: A Methodology for the Retrieve Stage of a Cardiovascular Case-Based Reasoning System** ..... 399  
Ana Duarte and Orlando Belo

**Machine Learning and Deep Learning**

**Forming of Validation Dataset for Deep Learning Based Model of Medical Image Grouping** ..... 411  
Robert Badarić, Franko Hrić, Mateja Napravnik, and Ivan Œtajduhar

**Deep Learning Based Radiomics to Predict Treatment Response Using Multi-datasets** ..... 431  
Thibaud Brochet, Jrme Lapuyade-Lahorgue, Alexandre Huat, Sbastien Thureau, David Pasquier, Isabelle Gardin, Romain Modzelewski, David Gibon, Juliette Thariat, Vincent Grgoire, Pierre Vera, and Su Ruan

**Convolutional Neural Network Classification of Liver Fibrosis Stages Using Ultrasonic Images Colorized by Features of Echo-Envelope Statistics** ..... 441  
Akiho Isshiki, Dar-In Tai, Po-Hsiang Tsui, Kenji Yoshida, Tadashi Yamaguchi, and Shinnosuke Hirata

**FedRNN: Federated Learning with RNN-Based Aggregation on Pancreas Segmentation** ..... 453  
 Zengtian Deng, Touseef Ahmad Qureshi, Sehrish Javed, Lixia Wang, Anthony G. Christodoulou, Yibin Xie, Srinavas Gaddam, Stephen Jacob Pandol, and Debiao Li

**UNet-2022: Exploring Dynamics in Non-isomorphic Architecture** ..... 465  
 Jiansen Guo, Hong-Yu Zhou, Liansheng Wang, and Yizhou Yu

**Hybrid-Fusion Transformer for Multisequence MRI** ..... 477  
 Jihoon Cho and Jinah Park

**STResNet: Covid-19 Detection by ResNet Transfer Learning and Stochastic Pooling** ..... 489  
 Wei Wang, Shui-Hua Wang, and Yu-Dong Zhang

**Convolutional Neural Networks for Newborn Pain Assessment Using Face Images: A Quantitative and Qualitative Comparison** ..... 503  
 Gabriel A. S. Coutrin, Lucas P. Carlini, Leonardo A. Ferreira, Tatianny M. Heiderich, Rita C. X. Balda, Marina C. M. Barros, Ruth Guinsburg, and Carlos E. Thomaz

**Machine Learning for the Evaluation and Detection of Key Markers in Dilated Cardiomyopathy** ..... 515  
 Xiaodan Bi, Zhenrun Zhan, Jinpeng Yang, Xu Tang, and Tingting Zhao

**Others**

**Schema Based Knowledge Graph for Clinical Knowledge Representation from Structured and Un-structured Oncology Data** .... 529  
 Farina Tariq, Saad Ahmad Khan, and Muhammad Moazam Fraz

**Intelligent Fuzzy Clinical Decision Support System to Classify Breast Cancer—Case Study: The Wisconsin Dataset** ..... 541  
 Y. F. Hernández-Julio, L. A. Díaz-Pertuz, M. Prieto-Guevara, M. Avilés-Román, B. Castillo-Osorio, M. Barrios-Barrios, and W. Nieto-Bernal

**Research on the Design and Production of VR Rehabilitation Game for Parkinson’s Disease Patients Based on Real-Time Action Acquisition** ..... 551  
 Ying Zhang, Xin Su, and Xibin Xu

**Force-Directed Graph Layout Based on Community Discovery and Clustering Optimization** ..... 561  
 Linshan Han, Beilei Wang, and Songyao Wang

**Comprehensive Strategy to Screen the Ankylosing Spondylitis-Related Biomarkers in the Peripheral Serum** ..... 573  
 Zhenrun Zhan, Xiaodan Bi, Xu Tang, and Tingting Zhao

# Medical Imaging



# Optimizing the Non-local Means Filtering of CT Images



Ivo Draganov  and Veska Gancheva

**Abstract** In this paper a general optimizing procedure is proposed for the non-local means (NLM) filter. It involves finding the optimal degree of smoothing, the size of the search window and the size of the comparison window for a series of Computed Tomography (CT) images. All of them contain Additive White Gaussian Noise (AWGN) with a particular variance and zero mean, both of which are preliminary unknown. Applying the optimization procedure over a single slice from the CT packet appears to be efficient enough in finding the optimal parameters of the filter for the rest of the CT images. Positive results are obtained from filtering a complete set of CT images from a patient's body and the quality of the filtration is higher than that of the Gaussian and Average filters.

**Keywords** CT image · Additive White Gaussian Noise · Non-local means filter · Optimization

## 1 Introduction

Computed Tomography (CT) images play a crucial role in medical diagnostics. Their quality is a prerequisite for effective medical treatment and it should be maintained as high as possible. The inherent noises from the principle of operation of the scanners worsen the overall representation of the internal organs, both in their homogenous areas and around the contours. Effective CT filtration could be established only if the involved filtering techniques preserve the structure of the organs as a whole.

In [1] Zhang et al. propose adaptive non-local means (NLM) filter which uses local principle neighborhoods (PC-NLM). Thus, they retain the structures of the organs from low-dose computed tomography (LDCT) images. The latter are known

---

I. Draganov (✉) · V. Gancheva  
Technical University of Sofia, 8 Kliment Ohridski Blvd., Sofia 1756, Bulgaria  
e-mail: [idraganov@tu-sofia.bg](mailto:idraganov@tu-sofia.bg)

V. Gancheva  
e-mail: [vgan@tu-sofia.bg](mailto:vgan@tu-sofia.bg)

to have significant level of noise and artifacts although the body of the patient is less affected by the radiation. The difference with the classical NLM filter is that Principal Component Analysis (PCA) is initially run over the local windows so they become decomposed to principal components. They are processed then by the NLM filter. Adaptive estimation of the filtering parameter is also proposed so components with higher Signal-to-Noise Ratio (SNR) are less changed than those with lower SNR. This preserves the original structures in the image. The whole procedure happens several times over typical LDCT images. The resulting Root Mean Square Error (RMSE) from tests for PC-NLM is 10.35 while for NLM it is 13.53, the Correlation Coefficient (CC) for PC-NLM is 0.9668, for NLM—0.8796, and the Structural Similarity Index Measure (SSIM) for PC-NLM is 0.7551, for NLM—0.5191.

In another study [2], Zhang et al. suggest the combination of Tensor Decomposition and Non-Local Means (TDNLM) for decreasing the extremely high levels of noise in spectral CT. The image projections from all energy channels are grouped together, forming a new image with higher SNR. Parameter selection strategy for the proposed approach is developed in order to get optimal quality of the images. Experimental results show decrease of RMSE from 0.225 to 0.0217  $\text{cm}^{-1}$  and increase of SSIM from 0.633 to 0.987.

Chen et al. [3] developed high-definition neural visualization technique of rodent brain. The authors use micro-CT scanning and the non-local means approach. This combination is thought to be effective in phenotyping and for histological manipulations. The NLM filter is applied as post-acquisition phase after the postnatal rat brain micro-CT scans for both the ex vivo and in vivo methods. The ex vivo method and the NLM filtering lead to 3D images close in details to  $4 \times$  light micrographs. This method provides more details in the neural features than those from the in vivo approach. On the other hand, the effect of the NLM filter on the in vivo samples is more underlined. It has bigger increase of the SNR. Resolutions of  $< 2\text{--}3 \mu\text{m}/\text{voxel}$  and scanning time  $> 15 \text{ h}$  are thought suitable to get satisfactory SNR.

Multi-scale transform and NLM is used as denoising approach for Positron Emission Tomography (PET) in [4] by Bal et al. It turns out that the mutual application of these two techniques preserves better both the isotropic and anisotropic components of the image rather than the application of just one processing algorithm. Wavelet and curvelet transform with Tree clustering NLM (TNLM) appears to be appropriate solution. TNLM takes out the homogenous (isotropic) features while the curvelet transform separates the edges and contours (anisotropic features). Filtration takes place separately for these two sets of objects. At the end they are grouped up together again. Positive results are reported from test with this approach with the additional benefit of increasing the contrast of filtered images.

Al-antari et al. [5] use the NLM filter for denoising high and low energy images obtained from Dual Energy X-ray Absorptiometry (DEXA). They adapted the filter parameters from uniform phantoms. The noises present in the source and the detector of the apparatus are modeled separately. SNR for high and low phantom images increases with 30.36% and 27.02%, respectively. In the same time, tests with real images of a spine reveal improvement of the SNR of 22.28 and 33.43%.

Another approach that employs multi-scale transform and the NLM filter for denoising PET images, in this instance dynamic ones, is proposed by Jomaa et al. [6]. Analyzing images of small animal hearts the authors take into account the correlation in time among them, employing the Shearlet and wavelet transforms to reduce the noise. Having noise level of 7.68% the chi-square parameter from the filtering is 4.06. Significant improvement in terms of the Peak Signal to Noise Ratio (PSNR) equal to  $74.38 \mp 9.2$  and SSIM is also achieved as well as in contrast –  $27.04 \mp 12.1$ .

A study on a wider set of images that aims denoising is described in [7] by Panigrahi et al. In the base of the approach is multiscale NLM filtering using curvelet transform and hard thresholding. In this case, ringing artefacts appear so additional processing by guided filter needs to be applied. Thus, edges and textures could be preserved better. The PSNR of the reconstructed images for noise with standard deviation  $\sigma = 40$  is 29.089 dB. For the NLM alone it is 27.252 dB. In the same time, SSIM is 0.777 for the composite technique and 0.691—for the NLM filter.

CT thoractic images are also being denoised by the NLM approach in a fast implementation (FNLM) as described by Kim et al. [8]. Gaussian noise with standard deviation of 0.002 is added to MASH phantom images and then filtered separately by the FNLM, Gaussian, median and Wiener filters. Achieved PSNRs are 82.354, 79.537, 82.094 and 81.882 dB. The Contrast to Noise Ratios (CNR) are 236.635, 47.630, 50.527 and 67.125, respectively.

FNLM is also used in the detection of pulmonary nodules as proposed by Shim et al. [9]. The processing is done over chest CT images. The  $h$  value of the filter is set to 2 values—0.0001 and 0.001. In the first case the registered number of artifacts is less. The Coefficient of Variation (COV) reaches in the best use case just above 2.5 and the Contrast to Noise Ratio (CNR)—around 22.

An adaptive NLM implementation is tried over Basis Material Images obtained from dual-energy CT at low emitting doses [10]. In this version of the filter distribution map helps in obtaining proper weights of the averaging pixels taking into account the decomposition error. The parameters of the filter in one of the experiments are  $h = 0.02$ ,  $r = 50$  and the radiuses of the search and comparison windows are 5 and 2, respectively. Depending on the type of the basis material and the dose level the PSNR changes between 20.35 and 31.61.

Obviously, there is large variety of implementations of the NLM filter and the different combinations with other techniques. In most of the cases, they are adapted for particular purposes. The main goal in this study is to propose a general scheme for optimizing the main parameters of the filter over CT image sets. It will lead to optimal results in the filtration process. The rest of the paper is organized as follows—in Sect. 2 description of the algorithm is given, in Sect. 3—experimental results and in Sect. 4—discussion, followed by a conclusion in Sect. 5.

## 2 Algorithm Description

### 2.1 The Non-local Means Filter

Let us have a grayscale noisy image  $I_n(i, j)$  with spatial coordinates of a pixel  $i \in \{0, M-1\}$  and  $j \in \{0, N-1\}$ . The dynamic range is  $I_n \in \{0, I_{max}\}$ . The filtered images are found according to [11]:

$$I_f(i, j) = \sum_k \sum_l w(i-k, j-l) I_n(i, j), \quad (1)$$

where  $k$  and  $l$  are such that  $(i-k) \in \{0, M-1\}$  and  $(j-l) \in \{0, N-1\}$  for all possible positions within the image;  $w(i-k, j-l)$  are weights which are estimated based on the similarity between neighborhoods around the pixels  $I_n(i, j)$  and  $I_n(i-k, j-l)$ . It is true that [11]:

$$\begin{cases} 0 \leq w(i-k, j-l) \leq 1 \\ \sum_k \sum_l w(i-k, j-l) = 1 \end{cases} \quad (2)$$

The weights themselves could be estimated from [11]:

$$w(i-k, j-l) = \frac{1}{C(i, j)} e^{-\frac{\sum_{p=-c/2}^{c/2} \sum_{q=-c/2}^{c/2} [I_n(i-p, j-q) - I_n(i-k-p, j-l-q)]_w^2}{h^2}}, \quad (3)$$

where  $p$  and  $q$  are temporal variables. They change within the boundaries of a comparison window with a size of  $c \times c$  pixels around the  $I_n(i, j)$  and  $I_n(i-k, j-l)$  pixels. Also,  $h$  is degree of filtering. As a lower index in the exponent nominator  $w$  denotes that it is a weighted Euclidean distance by a Gaussian with a fixed standard deviation of  $d$ . The parameter  $C(i, j)$  is coefficient of normalization, estimated as [11]:

$$C(i, j) = \sum_k \sum_l e^{-\frac{\sum_{p=-c/2}^{c/2} \sum_{q=-c/2}^{c/2} [I_n(i-p, j-q) - I_n(i-k-p, j-l-q)]_w^2}{h^2}}. \quad (4)$$

It could be shown that [11]:

$$\begin{aligned} & \frac{1}{K} \sum_{k=1}^K \sum_{p=-\frac{c}{2}}^{\frac{c}{2}} \sum_{q=-\frac{c}{2}}^{\frac{c}{2}} [I_n(i-p, j-q) - I_n(i-k-p, j-l-q)]_w^2 \\ &= \sum_{p=-\frac{c}{2}}^{\frac{c}{2}} \sum_{q=-\frac{c}{2}}^{\frac{c}{2}} [I_f(i-p, j-q) - I_f(i-k-p, j-l-q)]_w^2 + 2\sigma^2 \end{aligned} \quad (5)$$

where  $K$  is the number of comparisons between each two pairs of windows during the process of estimation of the similarity;  $\sigma$ —standard deviation of the noise in the image.

The comparisons are made within a search window with a bigger size of  $s \times s$  pixels. The third parameter that controls the filtration process is the Degree of Smoothing (DoS).

## 2.2 Proposed Optimization Procedure

The optimization procedure is shown in Fig. 1. One input image  $I(i, j)$  is filled with Additive White Gaussian Noise (AWGN) with variance  $\sigma$  and zero mean. Thus, we get the noisy image  $I_n(i, j)$ . All three control parameters of the filter DoS,  $s$  and  $c$  are varied in growing order with steps 1000, 2 and 2, respectively. There are 3 embedded loops for the purpose. In each iteration the PSNR and SSIM of the filtered image  $I_f(i, j)$  are calculated. After termination of all loops the maximal PSNR and SSIM determine the optimal  $\text{DoS}_{opt}$ ,  $s_{opt}$  and  $c_{opt}$ . Then the actual filtration of all the images from the CT set could take place.

The computational complexity of the proposed procedure with non-optimized version of the NLM filter is  $O(T((2c + 1)^2(2s + 1)^2 \cdot N \cdot M))$ . In the last expression  $T$  is the total number of iterations from all the loops shown in Fig. 1. It is straightforward to obtain such a relation given the considerations from [12].

## 3 Experimental Results

The test image set is excerpt of the DeepLesion dataset [13] and it is comprised of 103 CT images with dimensions  $512 \times 512$  pixels taken at 16 bpp bitdepth. The testing platform is IBM® PC® compatible computer with Intel® Core™ i7-6820HQ CPU with 4 cores. They are running in hyperthreading mode at 2.70 GHz, 64 GB of RAM and 1 TB HDD. The operating system is 64-bit MS® Windows® 10 Professional. The test environment is Matlab R2022a.

The first experiment aims to determine the optimal value of the Degree of Smoothing (DoS) which is varied between 1 and 65,535 with a step of 1000. The processing is done over a single image which is being noised with Additive White Gaussian Noise (AWGN) with a variance  $\sigma^2 = 0.01$  and zero mean. The resulting PSNR and SSIM of the reconstructed image is given in Fig. 2. Both curves change in such a way that saturation of the maximum value is reached at  $\text{DoS} = 19,000$ . It is considered as the optimum.

The processing time of the whole image is changing according to the curve from Fig. 3 with an average value of 0.1926s.

The second experiment consists of changing the size of the search window  $s$  and the size of the comparison window  $c$  using the optimal value of the DoS as a

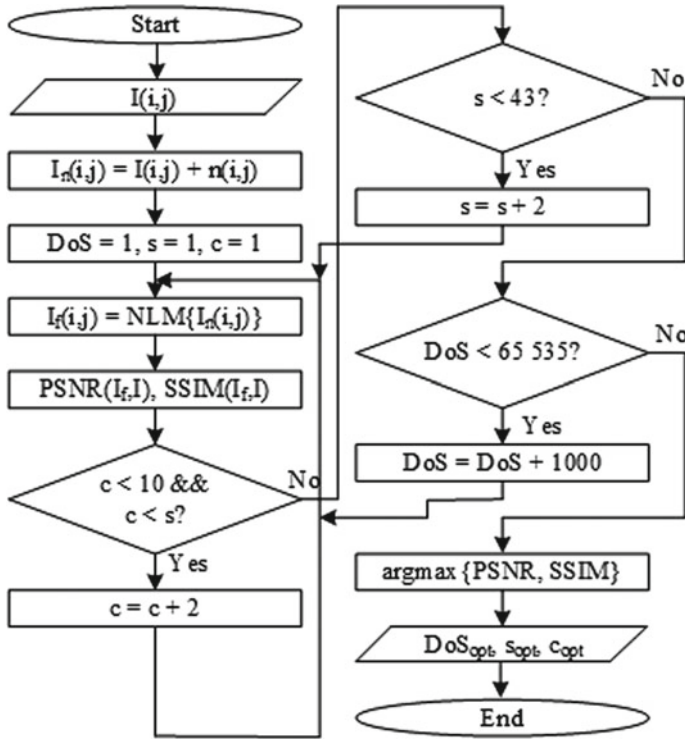


Fig. 1 Non-local means filter optimization procedure

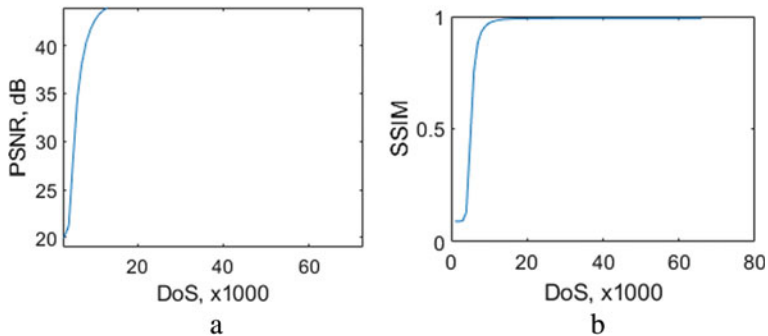
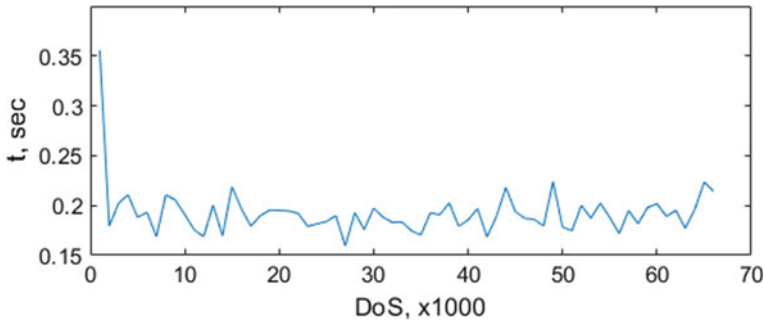
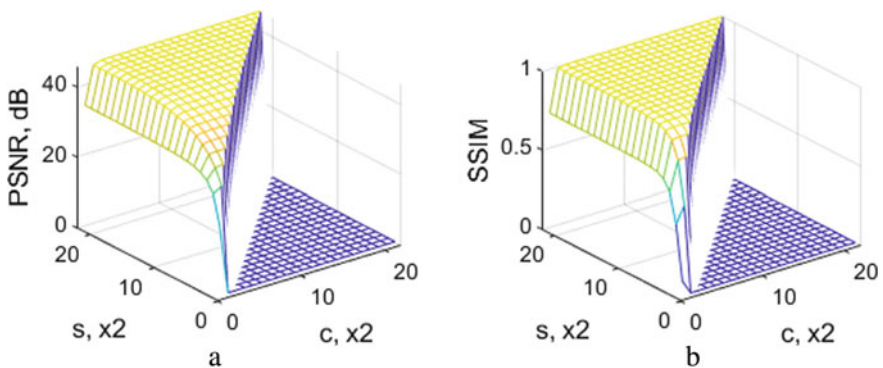


Fig. 2 Variation of **a** PSNR and **b** SSIM at different DoS

constant. The range for  $s$  is from 1 to 43. It is double the size of the typical value for this parameter as recommended in [14]. For each iteration with regards to  $s$  the size of the comparison window  $c$  changes from 1 to  $s - 1$ . Both windows' sizes are



**Fig. 3** Processing time of the NLM filter at different DoS

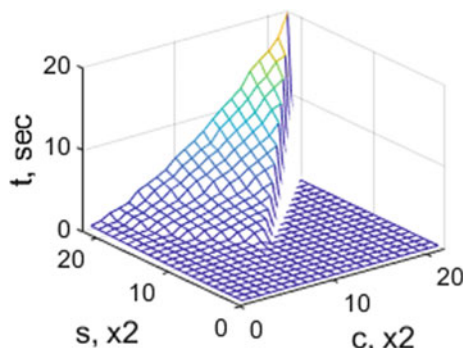


**Fig. 4** Variation of **a** PSNR and **b** SSIM at different  $s$  and  $c$

always an odd number. The resulting PSNR and SSIM are given in Fig. 4. These two parameters saturate to a maximum for  $c = 5$  and  $s = 43$ , which are the optimums.

The filtering time for all tested sizes  $s$  and  $c$  are shown in Fig. 5. It is monotonically rising function with the increase of both sizes.

The third experiment is related to filtering of all 103 CT images with  $\text{DoS}_{\text{opt}} = 19,000$ ,  $s_{\text{opt}} = 43$  and  $c_{\text{opt}} = 5$ . First AWGN with  $\sigma^2 = 0.001, 0.01$  and  $0.1$  is added to the images. Apart from the NLM filter, a Gaussian filter with zero centered kernel and corresponding to the noise standard deviation  $\sigma$  is also used. Together with it, an Average filter with size of the kernel of  $3 \times 3$  pixels is also applied over all 103 images. The PSNR and SSIM of the reconstructed images as well as the execution time in each case are presented in Table 1.



**Fig. 5** Filtering times at different  $s$  and  $c$

**Table 1** Efficiency of the applied filters

Parameter		Filter		
		NLM	Gaussian	Average
$\sigma^2 = 0.001$	PSNR, dB	47.08	30.00	39.47
	SSIM	0.9954	0.4941	0.9149
	t, sec	1.2953	0.0028	0.0013
$\sigma^2 = 0.01$	PSNR, dB	45.46	20.00	29.52
	SSIM	0.9939	0.0914	0.5374
	t, sec	1.2899	0.0016	0.0014
$\sigma^2 = 0.1$	PSNR, dB	32.21	11.18	20.44
	SSIM	0.6555	0.0127	0.1316
	t, sec	1.2923	0.0017	0.0014

## 4 Discussion

The highest value of the PSNR from the reconstructed images for noise variance 0.001 is obtained by the NLM filter. It is more than 7.6 dB than that of the Average filter and 17 dB difference with the Gaussian filter. The difference in SSIM between the NLM and the Average filter is relatively small—0.0805. It is considerably larger with the Gaussian filter—0.5013. However, the execution time of the NLM filter is 462.6 times higher than that of the Gaussian filter and 996.4 times the processing time of the Average filter. The latter is the fastest. These times remain almost constant regardless of the noise variance.

For higher levels of the noise— $\sigma^2 = 0.01$  the differences in PSNR between the NLM filter and the Gaussian and Average filters is 25.46 and 15.94 dB, respectively. These differences show increase in the efficiency of the NLM filter with the increase of the noise level, compared to the other two filters. For the most degraded images at



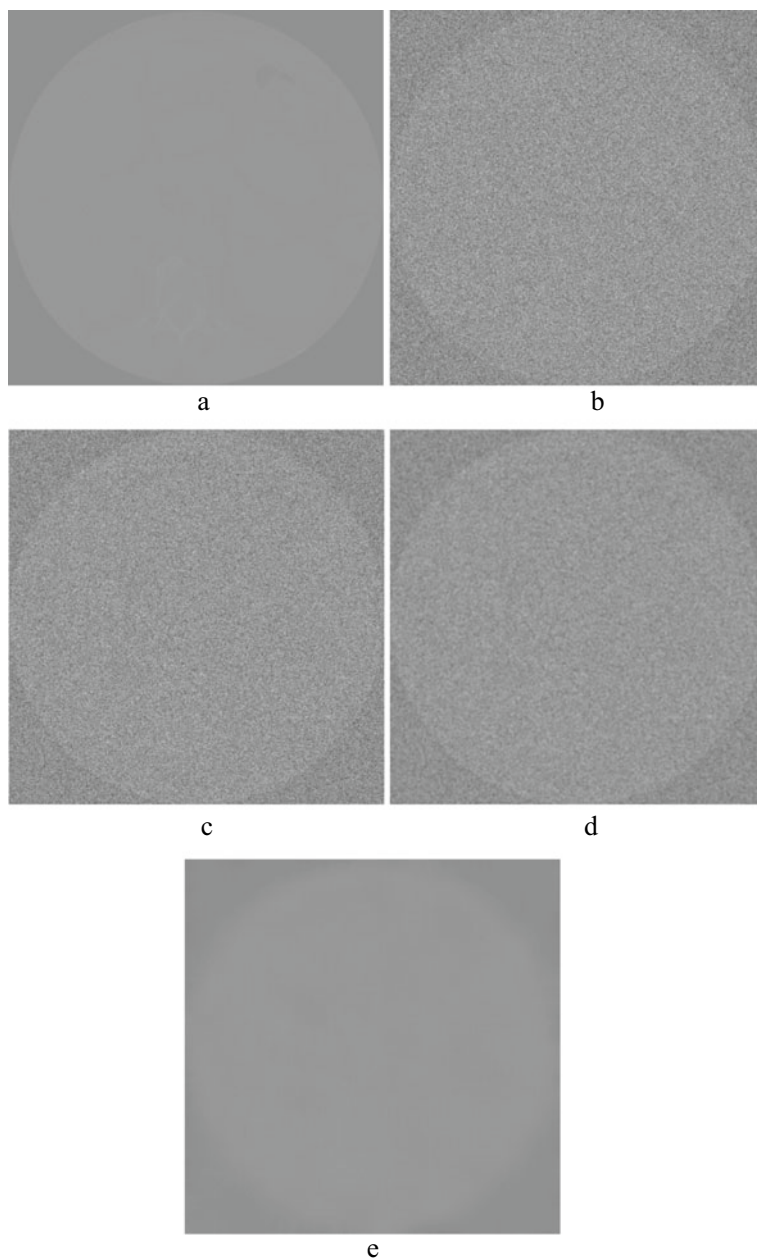
$\sigma^2 = 0.1$  the NLM filter outruns the Gaussian and Average filters in PSNR by 21.03 and 11.77 dB. For the SSIM the differences are 0.6428 and 0.5239, respectively.

The visual comparison of the quality of the reconstructed images show relatively small suppression of the noise for the Gaussian filter (Fig. 6c). In this instance the grainy structure over the whole image still persist. However, there is a little improvement in contrast and the structure of the organs could be spotted better. The Average filter (Fig. 6d) smooths the image more than the Gaussian filter. The contours of the objects are blurred and the size of the grains, still remaining from the noise, is bigger. The smoothest effect from the filtering comes from the NLM filter (Fig. 6d). The grainy structure is totally absent, but the contrast is a bit lower.

The quality of the NLM filtered images depends non-linearly from the DoS (Fig. 2a and b). After steep increase of both the PSNR and SSIM from values of 20 dB and 0.1 for DoS = 1, there is a zone of saturation starting around DoS = 19,000. There the PSNR is around 44 dB and the SSIM reaches almost 1. The filtering time does not seem to depend on the DoS. There is just a slight variation for it within the interval 0.15–0.22 s (Fig. 3). The change of the PSNR is upwards with the increase of the search window with its side  $s$  reaching a maximum for 43 pixels of around 45.5 dB. From  $s = 3$  with PSNR = 27.81 dB to  $s = 21$  with PSNR = 33.86 dB there is the most significant increase interval. In the interval  $c = 3$  with PSNR = 29.43 dB up to  $c = 11$  with PSNR = 40.26 dB the quality of images rises the most entering a saturation zone which ends at  $c = 43$  with PSNR = 45.55 dB. SSIM almost identically follows the change of the PSNR with the lowest level of 0.5347 for  $c = 3$  and  $s = 3$ . Then it goes to 0.9451 for  $s = 9$  and  $c = 3$  and then follows the saturation zone with 0.9947 for  $s = 43$  and  $c = 41$  at its end. Filtering time steadily increases with the growth of  $s$  and  $c$  (Fig. 5). From  $5 \times 5$  pixels search window and  $3 \times 3$  comparison window it is 0.0057 s and rises to 19.68 s for  $s = 43$  and  $c = 41$ .

## 5 Conclusions

In this paper a general optimization scheme is proposed for the control parameters of the NLM filter. It is tested over a set of CT images. Experimental results show that the Degree of Smoothing affects the quality of the reconstructed images. The increase of this parameter leads to saturation of both the PSNR and SSIM. There is a minimal value for DoS which could be found as an optimal at the beginning of the saturation zone. The change of DoS has no significant effect on the filtration time. The sizes of the search and comparison windows also have non-linear effect over the quality of the reconstructed images. For both of them there are saturation areas in the PSNR and SSIM functions. It is possible to select the minimum windows sizes, such that they lay at the beginning of the saturation zone. Thus, they guarantee best quality of the images at the lowest computational time. The computational time, itself, increases monotonically with the increase of the surface of the search and comparison window. The NLM filter provides better quality of the filtered CT images than the Gaussian and Average filters for wide range of noise level of AWGN. The filtering time of all



**Fig. 6** Sample CT image: **a** original, **b** noisy, filtered by **c** Gaussian filter, **d** average filter and **e** optimal NLM filter (the representation of the images here is at 8 bpp, scaled down from 16 bpp)

three filters does not depend on the noise level. The NLM filter is more than 2 orders of a magnitude slower than the other two filters. There is no grainy structure in the images, filtered by the NLM, but there is a little loss of contrast. As a future work optimization of the NLM filter as processing time could be undertaken. Also testing with other types of images could be accomplished, e.g. magneto-resonance imaging (MRI), multispectral and hyperspectral.

## References

1. Zhang, Y., Lu, H., Rong, J., Meng, J., Shang, J., Ren, P., Zhang, J. Adaptive non-local means on local principle neighborhood for noise/artifacts reduction in low-dose CT images. *Medical Physics* 44(9), e230–e241 (2017).
2. Zhang, Y., Salehjahromi, M., Yu, H. Tensor decomposition and non-local means based spectral CT image denoising. *Journal of X-ray Science and Technology* 27(3), 397–416 (2019).
3. Chen, K. C., Arad, A., Song, Z. M., Croaker, D. High-definition neural visualization of rodent brain using micro-CT scanning and non-local-means processing. *BMC medical imaging* 18(1), 1–13 (2018).
4. Bal, A., Banerjee, M., Chaki, R., Sharma, P. An efficient method for PET image denoising by combining multi-scale transform and non-local means. *Multimedia Tools and Applications* 79(39), 29087–29120 (2020).
5. Al-antari, M. A., Al-masni, M. A., Metwally, M., Hussain, D., Valarezo, E., Rivera, P., Gi, G., Park, J. M., Kim, T. Y., Park, S.-J., Shin, J.-S., Han, S.-M., Kim, T. S. Non-local means filter denoising for DEXA images. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 572–575. IEEE (July 2017).
6. Jomaa, H., Mabrouk, R., Khelifa, N., Morain-Nicolier, F. Denoising of dynamic PET images using a multi-scale transform and non-local means filter. *Biomedical Signal Processing and Control* 41, 69–80 (2018).
7. Panigrahi, S. K., Gupta, S., Sahu, P. K. Curvelet-based multiscale denoising using non-local means & guided image filter. *IET Image Processing* 12(6), 909–918 (2018).
8. Kim, B. G., Kang, S. H., Park, C. R., Jeong, H. W., Lee, Y. Noise level and similarity analysis for computed tomographic thoracic image with fast non-local means denoising algorithm. *Applied Sciences* 10(21), 7455 (2020).
9. Shim, J., Yoon, M., Lee, M. J., Lee, Y. Utility of fast non-local means (FNLM) filter for detection of pulmonary nodules in chest CT for pediatric patient. *Physica Medica* 81, 52–59 (2021).
10. Yuan, Y., Zhang, Y., Yu, H. Adaptive non-local means method for denoising basis material images from dual-energy CT. *Journal of Computer Assisted Tomography* 42(6), 972 (2018).
11. Buades, A., Coll, B., Morel, J. M. A non-local algorithm for image denoising. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 2, pp. 60–65. IEEE (June 2005).
12. Condat, L. A simple trick to speed up the non-local means. hal-00512801, version, 1 (2010).
13. Ke Yan, Xiaosong Wang, Le Lu, Ronald M. Summers. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging* (2018). <https://doi.org/https://doi.org/10.1117/1.JMI.5.3.036501>
14. Mathworks, Non-local means filtering of image, <https://www.mathworks.com/help/images/ref/imnlmfilt.html>, last accessed on August 4th, 2022.

# Exploring Structure-Wise Uncertainty for 3D Medical Image Segmentation



Anton Vasiliuk, Daria Frolova, Mikhail Belyaev, and Boris Shirokikh

**Abstract** When applying a Deep Learning model to medical images, it is crucial to estimate the model uncertainty. Voxel-wise uncertainty is a useful visual marker for human experts and could be used to improve the model's voxel-wise output, such as segmentation. Moreover, uncertainty provides a solid foundation for out-of-distribution (OOD) detection, improving the model performance on the image-wise level. However, one of the frequent tasks in medical imaging is the segmentation of distinct, local structures such as tumors or lesions. Here, the structure-wise uncertainty allows more precise operations than image-wise and more semantic-aware than voxel-wise. The way to produce uncertainty for individual structures remains poorly explored. We propose a framework to measure the structure-wise uncertainty and evaluate the impact of OOD data on the model performance. Thus, we identify the best UE method to improve the segmentation quality. The proposed framework is tested on three datasets with the tumor segmentation task: LIDC-IDRI, LiTS, and a private one with multiple brain metastases cases.

**Keywords** Uncertainty estimation · Out-of-distribution detection · Segmentation · CT · MRI

## 1 Introduction

Advances in Deep Learning (DL) allow solving a medical image segmentation task with near human-level quality [1]. But predictions of DL models in medical imaging could not be taken blindly and assumed to be accurate. Ideally, the model is required to provide the uncertainty estimate of its output. Estimating uncertainty maps in medical

---

A. Vasiliuk · D. Frolova · M. Belyaev · B. Shirokikh  
Artificial Intelligence Research Institute (AIRI), Moscow, Russia

A. Vasiliuk  
Moscow Institute of Physics and Technology, Moscow, Russia

D. Frolova · M. Belyaev · B. Shirokikh (✉)  
Skolkovo Institute of Science and Technology, Moscow, Russia  
e-mail: [boris.shirokikh@skoltech.ru](mailto:boris.shirokikh@skoltech.ru)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
R. Su et al. (eds.), *Medical Imaging and Computer-Aided Diagnosis*, Lecture Notes in Electrical Engineering 810, [https://doi.org/10.1007/978-981-16-6775-6\\_2](https://doi.org/10.1007/978-981-16-6775-6_2)

image segmentation helps to solve a wide range of problems. The uncertainties are desired for a better reception by medical experts [2], but the direct impact is hard to measure in this case. Alternatively, one uses uncertainty on a voxel-wise level to refine the segmentation map, thus improving the model’s performance [3]. Uncertainty maps also could be aggregated on an image-wise level, forming a standalone out-of-distribution (OOD) detection method [4].

In the case of multiple objects or *structures* per image (e.g., tumors, lesions), clinical tasks also require analyzing the model’s output on the structure-wise level. Such cases are common in medical, especially radiological [5], imaging: a brain tumor, lung cancer, organ-at-risk, or liver tumor segmentation. However, the ways of using or reporting the uncertainty on distinctly localized multiple structures are poorly explored, rising acute questions. For example, using voxel-wise uncertainty, as in [3], one can improve the segmentation quality of detected structures but cannot filter individual false positive (FP) predicted objects. In image-wise uncertainty, as in [4], we do not consider the segmentation of local structures and also rebalance FP and true positive (TP) predictions in a sub-optimal way, filtering the whole image at once.

Therefore, in this work, we study uncertainty for individual predicted structures, i.e., connected areas of the predicted segmentation mask. We assume that treating uncertainty maps in a structure-wise manner helps to remove the FP detections more effectively, thus improving the detection quality. Secondly, we assume that structure-wise uncertainty (SWU) value strongly correlates with the segmentation quality of a given structure. If the latter assumption holds true, it’s possible to conduct quality control to enhance the model segmentation performance in the human-in-the-loop setup [6], where the human expert refines the most uncertain (thus, worst) predictions. We validate and experimentally confirm both assumptions.

Partially, these assumptions were tested directly or indirectly in a prior work. We detail the related studies and compare with their methodology in Sect. 2. We extend these studies in several major ways and below we detail our contributions:

- *Structure-wise uncertainty estimation.* We evaluate different uncertainty estimation (UE) techniques and local uncertainty aggregation functions. We show that switching from predicted values space to the structure’s Entropy produces 3% fewer FP predictions on average, up to 7% fewer on LiTS dataset, adding a negligible overhead and being applicable to any segmentation network.
- *Uncertainty under out-of-distribution.* We propose to evaluate aleatoric and epistemic performance by testing on in-distribution (ID) and OOD data. We develop three OOD aleatoric setups to demonstrate different SWU properties. We show that Pairwise-Dice Uncertainty [7] excels in the OOD setups, filtering out 6% more FP predictions than the baseline method, and itself in the ID setups.
- *Extensive and robust evaluation.* We compare state-of-the-art UE techniques on three large datasets with volumetric medical images. The datasets relate to the described problem and contain cases with multiple lesions.

## 2 Related Work

One of the direct SWU applications is FP reduction. Pursuing this goal, Nair et al. [8] improved performance of their model in the multiple sclerosis segmentation task. The authors took a sum of logarithms (*sum-log*) over a predicted structure uncertainty as a score to filter them. We argue that the sum-log is biased towards small objects. In a broader setup with the differently sized target structures, we show that the standard aggregation techniques such as *mean* surpass sum-log with a great margin.

Another approach to filter FP is a dedicated postprocessing model. Ozdemir et al. [9] trained a network to classify predicted structures and compared different dropout and ensembling regimes for this network. Bhat et al. [10] reduced FP in the liver lesion segmentation task by training an SVM classifier on predicted patches, their uncertainties, and hand-crafted features. However, FP reduction with a separate network is limited with strictly one structure per patch or image. Here, we consider a more general setup.

Other studies explore the ability to predict quality from uncertainty. Roy et al. [7] developed a Monte-Carlo-based approach to predict whole-brain segmentation and uncertainty maps. The authors calculated mean entropy, pairwise Dice score, coefficient of the volume variation, and intersection over union to predict structure-wise Dice scores. Mehrtash et al. [11] proposed to use mean entropy to predict structure Dice scores and achieved a high Pearson correlation between them for different tasks. Hoebel et al. [12] studied several setups for the whole image quality prediction. They compared Deep Ensembles against Monte-Carlo dropout and Dice loss against weighted cross-entropy in terms of pairwise Dice score, coefficient of volume variation, and mean entropy value. DeVries et al. [13] trained a separate network to predict image-wise segmentation quality and compared different uncertainty estimation methods with this network. We extend these approaches by studying uncertainty application in a structure-wise manner instead of the image-wise one and evaluate all related UE techniques. Moreover, we introduce studying uncertainty in the OOD setup.

SWU is also taken advantage of in other challenges. Seeböck et al. [14] developed an anomaly detection method for retinal optical coherence tomography, but the authors pursue the other goal of developing a weakly-supervised segmentation model. Hiasa et al. [15] studied muscle segmentation in an active learning setting and proposed to use mean structure-wise variance to predict the structure's Dice score. In our work, we identify the UE technique for the supervised segmentation problem.

Thereby, we conduct an extensive study of known uncertainty estimation techniques on a structure-wise level. We perform unified experiments across individual aggregation and uncertainty estimation techniques, emphasizing the importance of studying both aleatoric and epistemic setups.

### 3 Methods

In this section, we propose a general framework to estimate SWU. The estimation process consists of three steps: (i) compute a voxel-wise uncertainty map, (ii) split the segmentation map to obtain the individual structures, and (iii) aggregate the uncertainty inside every structure. The SWU scores can be further used for the FP filtration and quality estimation.

#### 3.1 Structure Definition

The ground truth structure (e.g., lesion, tumor) is defined as a connected area of the annotation mask. Similarly, a predicted structure is a connected area of the predicted segmentation mask, which can be binarized with different probability thresholds. We experimentally compared different threshold values and found out that either larger (e.g., 0.75) and smaller (e.g., 0.25) ones give considerably worse results than the de facto standard threshold of 0.5. We further use the probability threshold of 0.5 to define a predicted structure and omit the comparison of thresholds for the clarity.

#### 3.2 Uncertainty Estimation Methods

To obtain uncertainty maps, we use Deep Ensembles [16], which are considered to be state-of-the-art for estimating uncertainty in the medical image segmentation tasks [11, 17]. We construct an ensemble of  $T = 5$  neural networks trained with different weight initializations, during the inference time,  $T$  predicted probability maps  $P_1, \dots, P_T$  are generated for an input image. If probability map is a multi-channel (softmax) output, the different channels are denoted as  $P_t^c$ .

The conventional way to filter FP predictions is to threshold a predicted mask with its maximum value; thus, we consider **Pred (max)** a baseline method. We also use the output of the final layer before sigmoid activation instead of probabilities and call this method **Logit**. As one of the standard UE methods, we include **Entropy**:  $U_{\text{Ent}} = -\sum_{c=1}^C P^c \log P^c$ .

The methods above can be applied both to a single and the ensemble's (i.e., the average) prediction by substituting  $P^c$  with  $\bar{P}^c$ , where  $\bar{P}^c = \frac{1}{T} \sum_{t=1}^T P_t^c$ . Alternatively, we can apply averaging after calculating the entropy:  $U_{\text{AE}} = -\frac{1}{T} \sum_{t=1}^T \sum_{c=1}^C P_t^c \log P_t^c$ . We call this method **Average entropy (AE)** and also include it into consideration.

The following two methods are drawn from the related work on UE and operate only on multiple predicted probability maps. The first is **Mutual Information (MI)** or BALD [18]:  $U_{\text{MI}} = -\sum_{c=1}^C \bar{P}^c \log \bar{P}^c + \frac{1}{T} \sum_{t=1}^T \sum_{c=1}^C P_t^c \log P_t^c$ . The second is **Voxel-wise variance** of predictions [19]:  $U_{\text{Var}} = \frac{1}{TC} \sum_{t=1}^T \sum_{c=1}^C (P_t^c - \bar{P}^c)^2$ .

The last method that we consider is **Pairwise Dice (PD)** between predictions [7]. Unlike previous methods, it produces a single uncertainty score per structure instead of a voxel-wise uncertainty map. This uncertainty score is the averaged dice score between all pairs of  $T$  predictions and a given structure.

### 3.3 Uncertainty Aggregation Techniques

Assuming uncertainty map is given, we need to assign a single score for every structure. In [8], the authors calculate *sum-log* of uncertainties for all voxels  $v$  in the structure  $S$ :  $u = \sum_{v \in S} \log U_v$ . We argue that *sum-log* is heavily unbalanced in cases with differently sized structures, which are common. Therefore, we include in comparison the standard and, in this case, balanced statistics: *min*, *max*, *mean*, and *median*.

## 4 Experiments

### 4.1 Data

We study SWU performance on three different challenges. To explore a method’s aleatoric performance compared to epistemic, we provide an OOD dataset in every task. The model is trained only on the ID training set, and we compare its performance on the ID test set and the OOD data. All OOD datasets share the same preprocessing steps with their ID pairs; the preprocessing is disclosed in the supplementary materials.

**Mets** (*private* ID dataset) includes 1554 T1-weighted head MR images with annotated metastases masks. Besides, one may consider a recently published public alternative [20]. **EGD** (OOD for *Mets*) includes 374 images of brain MRI (4 different modalities) with annotated glioblastoma masks [21]. We select 141 of them with Flair as the primary modality. We consider it to have empty metastases masks. **LIDC** (ID) includes 1018 chest CT images from LIDC/IDRI database [22] with annotated lung cancer masks. **MIDRC** (OOD for *LIDC*) includes 110 chest CT images with annotated COVID-19 lesion masks [23]. We select 98 of them with non-empty segmentation masks. We consider it to have empty lung cancer masks. **LiTS** (ID) includes 131 abdominal CT images with annotated liver and liver tumor masks [24]. **LiTS-mod** (OOD for *LiTS*) is a synthetically created dataset from 13 *LiTS* images with empty liver tumor masks, generating typical CT imaging artifacts [25, 26].

All considered ID datasets are diverse and have the multiple small structures segmentation task, which satisfies the considered setup. Five out of six datasets are publicly available, yielding the partial reproducibility of our experiments.



## 4.2 Experimental Setup

In all our experiments, we use the same segmentation model based on nnU-Net [27]. The implementation and training details are provided in the supplementary materials and they are also available in our repository.<sup>1</sup>

*Metrics.* To measure FP reduction capacity, we evaluate how many FP detections per image are filtered at 95% recall level and compute average recall for high precision values. Considering  $R_{\max}$  is the maximum model’s recall value, and  $F_x$  is the average number of FP predictions for a recall value  $\frac{x}{100} \times R_{\max}$ , we compute the FP reduction metric as  $\frac{F_{100}-F_{95}}{F_{100}}$  to account for a different number of FP on the OOD setups. The average recall is computed for precision values  $P$  from  $\min(P)$  to  $\frac{1}{2}(\min(P) + \max(P))$ , the same for each method on a setup, to obtain statistics only from the more relevant high recall region. For quality control metrics, we report the absolute value of the Spearman correlation coefficient between the individual structure Dice scores and SWU values.

## 4.3 Results

**FP reduction.** Despite the solid performance of the baseline method, there are advantages of using other uncertainty measures and aggregation techniques; see Table 1. Using the Entropy measure or mean aggregation, one can produce fewer FP predictions for most setups. Except for Variance, AE, and *Entropy (sum-log)* [8], the other methods surpass the baseline. The most consistent methods are PD and Entropy, allowing for up to 7% FP reduction with a single model and 11% with the ensemble model.

A considerable rise in OOD performance is shown by the *discrepancy* methods (PD, MI, Variance) in comparison to the *averaging* methods (Pred, Logit, AE, Entropy); see Fig. 1, Table 1. The discrepancy methods produce 6–8% fewer FP predictions on the aleatoric OOD setups while averaging methods do not exceed 3% limit, with an even more apparent difference on individual datasets. Since the OOD setups differ from the ID ones only in the additional FP samples, we can conclude that the discrepancy methods are better at filtering OOD data.

**Quality control.** For most of the methods, *mean* aggregation is a better index of a structure quality than *min* and *max* aggregations (Fig. 2 and Table 2). The only exceptions are Variance, with poor performance in all setups, and Pairwise Dice, which does not use the voxel space uncertainty. The *Entropy (mean)* is the best method in all LIDC and Mets setups and the second best in ensemble LiTS setup, while discrepancy methods generally show the weaker correlation.

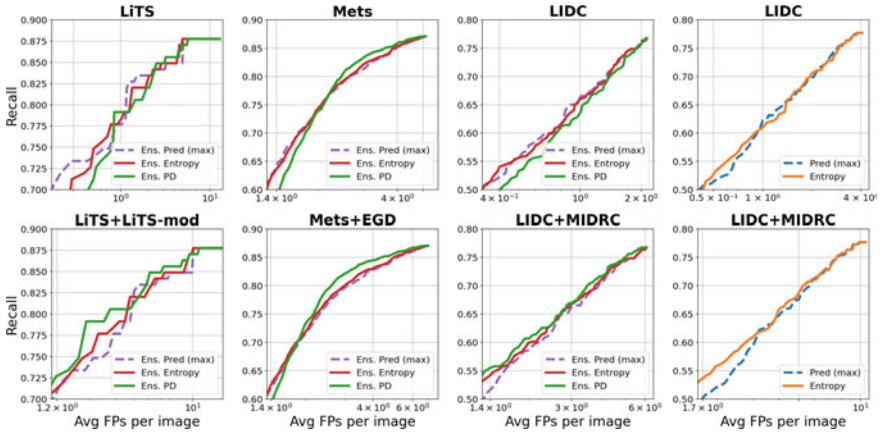
---

<sup>1</sup> <https://github.com/BorisShirokikh/u-froc>.

Table 1 FP reduction capabilities of different SWU methods

Method	Ensemble	FP reduction						Average recall						
		Alea.	LiTS	LiTS*	LIDC	LIDC*	Met	Met*	LiTS	LiTS*	LIDC	LIDC*	Mets	Mets*
Pred (max)		0.52	0.75	0.78	0.36	<b>0.33</b>	0.45	0.51	<b>0.86</b>	<b>0.84</b>	0.69	0.58	0.83	0.80
Pred		0.53	0.80	0.82	0.31	0.30	<b>0.48</b>	<b>0.56</b>	<b>0.86</b>	<b>0.84</b>	0.69	0.61	<b>0.84</b>	<b>0.81</b>
Logit		0.53	0.79	0.82	0.31	0.30	<b>0.48</b>	<b>0.56</b>	<b>0.86</b>	<b>0.84</b>	0.69	0.61	<b>0.84</b>	<b>0.81</b>
Entropy		<b>0.54</b>	<b>0.82</b>	<b>0.83</b>	0.32	0.32	<b>0.48</b>	<b>0.56</b>	<b>0.86</b>	<b>0.84</b>	0.69	<b>0.62</b>	<b>0.84</b>	<b>0.81</b>
Entropy (min)		0.52	0.77	0.79	0.36	<b>0.33</b>	0.44	0.51	<b>0.86</b>	<b>0.84</b>	0.69	0.59	0.83	0.80
Entropy (sumlog[8])		0.51	0.75	0.75	<b>0.37</b>	<b>0.33</b>	0.41	0.45	0.85	0.83	<b>0.70</b>	0.47	0.81	0.71
Pred (max)	✓	0.49	0.83	0.81	0.27	0.26	0.37	0.44	0.86	0.85	0.67	0.58	0.83	0.82
Pred	✓	0.49	0.82	0.81	0.27	0.29	0.39	0.47	<b>0.87</b>	0.85	0.67	0.60	0.83	0.83
Logit	✓	0.50	0.83	0.82	0.27	0.29	0.39	0.47	<b>0.87</b>	0.85	0.67	0.61	<b>0.84</b>	0.83
AE (min)	✓	0.44	0.85	0.83	0.22	0.13	0.26	0.28	0.86	0.84	0.67	0.55	0.82	0.81
Entropy	✓	0.50	0.84	0.82	<b>0.28</b>	0.29	0.39	0.47	<b>0.87</b>	0.85	0.67	0.61	0.83	<b>0.84</b>
Entropy (min)	✓	<b>0.51</b>	<b>0.89</b>	<b>0.85</b>	0.26	0.27	0.37	0.44	0.86	0.85	<b>0.68</b>	0.58	0.82	<b>0.84</b>
Entropy (sumlog[8])	✓	0.43	0.71	0.70	0.27	0.26	0.32	0.38	0.86	0.79	<b>0.68</b>	0.44	0.82	0.79
MI	✓	0.47	0.79	0.79	0.21	<b>0.32</b>	0.42	0.54	0.86	0.84	0.64	<b>0.63</b>	<b>0.84</b>	<b>0.84</b>
PD	✓	0.50	0.82	<b>0.83</b>	0.23	0.30	<b>0.44</b>	<b>0.55</b>	<b>0.87</b>	<b>0.86</b>	0.66	<b>0.63</b>	<b>0.84</b>	<b>0.84</b>
Variance (min)	✓	0.40	0.65	0.68	0.14	0.26	0.41	0.50	0.83	0.80	0.63	0.58	<b>0.84</b>	0.83

Metrics are described in Sect. 4.2. Datasets with added OOD counterpart data are marked by \*. Mean aggregation is used unless it is stated in brackets. Average FP reduction for ID and OOD setups is provided in the Epi. and Alea. columns, respectively



**Fig. 1** FROC curves for the best methods in comparison with the baseline (the dashed line). For visual clarity, the average number of FP per image is given in the log scale. The measures are obtained on the ID data (row 1) and the ID and OOD data combined (row 2). *Discrepancy methods show better performance when OOD dataset is present*

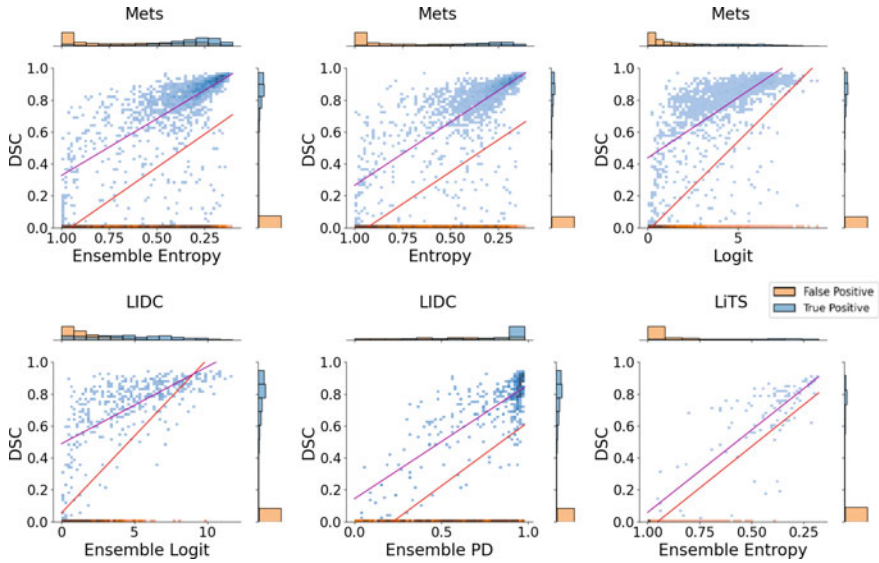
Note that we do not consider the OOD setups in this quality control study, since OOD data only introduces FP instances and, thus, does not affect the correlation of scores on TP instances.

Overall, the most consistent method to evaluate SWU is the mean Entropy. It performs among the top methods, producing 2.5% fewer FP predictions on average and giving a 0.77 Spearman correlation with the object Dice score for TP predictions. In the presence of OOD data, Pairwise Dice score reduces FP predictions better than others, filtering from 2 to 11% more FP structures, depending on the OOD setup.

#### 4.4 Discussion

To construct OOD setups with positive samples, we had to include the ID data. That means that the FP reduction metric shows an average between ID and OOD false positives, and pure OOD performance remains unknown. One of the possible ways to approach this problem is to create a domain-shifted setup which would contain OOD data with the true-positive structures.

The other promising application of the SWU framework is a more efficient human-in-the-loop control. Quality estimates might be a good measure to select images or individual structures to show a medical professional, but the question of how to gain the most quality given a limited amount of human interaction, combined with optimal FP reduction remains open.



**Fig. 2** Linear models for object Dice scores from SWU plotted on top of the structures' heat-maps. The purple and red lines are constructed for only TP and both TP and FP structures, respectively. Single-dimensional distributions of the measures are plotted along the axes

**Table 2** Spearman correlation coefficients between structure's Dice score and SWU value for TP predictions

	Second Agg.	Ensemble	LiTS	LiTS**	LIDC	LIDC**	Mets	Mets**
Pred	Max		<b>0.86</b> /0.81	0.46/0.46	0.68/0.62	0.63/0.63	<b>0.79</b> /0.65	0.61/0.60
Logit	Max			0.46/0.46	0.67/0.62	<b>0.64</b> /0.63	0.75/0.65	0.61/0.60
Entropy	Min		<b>0.86</b> /0.81	0.46/0.46	<b>0.69</b> /0.62	<b>0.64</b> /0.63	<b>0.79</b> /0.65	0.61/0.60
Pred	Max	✓	<b>0.79</b> /0.69	0.66/0.66	<b>0.70</b> /0.64	0.66/0.66	<b>0.78</b> /0.64	<b>0.61</b> /0.59
Logit	Max	✓	0.75/0.69	0.66/0.66	0.69/0.64	<b>0.67</b> /0.66	0.74/0.64	<b>0.61</b> /0.59
AE	Min	✓	0.72/0.69	0.55/0.65	0.55/0.61	0.58/0.65	0.77/0.64	0.53/0.58
Entropy	Min	✓	0.78/0.69	0.66/0.66	<b>0.70</b> /0.64	<b>0.67</b> /0.66	<b>0.78</b> /0.64	<b>0.61</b> /0.59
MI	Min	✓	0.72/0.24	0.64/0.48	0.49/0.38	0.48/0.40	0.57/0.27	0.50/0.33
PD	Min	✓	0.76/0.74	0.65/ <b>0.71</b>	0.67/0.66	0.64/0.63	0.74/0.76	0.59/0.60
Variance	Min	✓	0.42/0.64	0.44/0.60	0.22/0.55	0.22/0.55	0.10/0.52	0.22/0.54

Columns denoted by “\*\*” show values for all predictions, including FP. The values separated by “/” represent *mean* and extreme aggregation, respectively

## 5 Conclusion

In this work, we have conducted an extensive study of structure-wise uncertainty over six different setups. We have shown that mean Entropy provides a solid baseline in both false positive reduction and quality control tasks. Also, we have revealed

the importance of studying uncertainty metrics under different origins of data. In our experiments, the discrepancy SWU methods perform significantly better for FP reduction in the presence of the OOD data, with the best results achieved by Pairwise Dice. Provided results should serve as a solid baseline for future structure-based analysis.

**Acknowledgements** The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study. This research was funded by Russian Science Foundation grant number 20-71-10134.

## Experimental Setup

### *Preprocessing*

Here, we describe data preparation steps including datasets splits, normalization, and interpolation.

**Mets** data is randomly split into train (1140 images) and test (414 images) sets. We interpolate the images to have  $1\text{ mm} \times 1\text{ mm} \times 1\text{ mm}$  spacing.

**LIDC** data is randomly split into train (812 images) and test (204 images) sets. We clip image intensities between  $-1350$  and  $350$  Hounsfield units (HU)—the standard lung window. We interpolate images to have  $1\text{ mm} \times 1\text{ mm} \times 1.5\text{ mm}$  spacing.

**LiTS** is presented as two subsets, so we use the first as a test (28 images) and the second, excluding cases with empty tumor masks, as a train (90 images) set. The images are cropped to the provided liver masks. The intensities are clipped to the  $[-150, 250]$  HU interval—the standard liver window. Finally, we interpolate images to have  $0.77\text{ mm} \times 0.77\text{ mm} \times 1\text{ mm}$  spacing.

**LiTS-mod** is obtained by random changes of the reconstruction kernel to be extremely soft ( $a = -0.7, b = 0.5$ ) or sharp ( $a = 30, b = 3$ ) using the implementation and notations of [26], and addition of “metal” artifacts (ball of radius 5 and 3000 HU) by substituting the parts of sinogram projection, as in [25].

Before passing through the network, we scale image intensities in  $[0, 1]$ .

### *Training Setup*

Although using cross-entropy loss has theoretical justifications of encouraging better calibrated predictions [16], models trained with this loss function fail in our segmentation task. For that reason we use Dice Loss [28] and its modifications in our experiments. Thus, uncertainty estimates might be shifted in such tasks, and experimental evaluation, as in our study, becomes even more relevant. All models are trained in a patch-based manner: patches are sampled randomly so that they con-

tain structures. We use SGD optimizer with Nesterov momentum of 0.9 and  $10^{-3}$  initial learning rate, which is decreased to  $10^{-4}$  after 80% of epochs. For LiTS and Mets segmentation the model is trained for 100 epochs (100 iterations per epoch, batch size 20), while for LIDC segmentation there are 30 epochs (1000 iterations per epoch, batch size 2).

## References

1. Lee, J.G., Jun, S., Cho, Y.W., Lee, H., Kim, G.B., Seo, J.B., Kim, N.: Deep learning in medical imaging: general overview. *Korean journal of radiology* 18(4), 570–584 (2017)
2. Kompa, B., Snoek, J., Beam, A.L.: Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine* 4(1), 1–6 (2021)
3. Iwamoto, S., Raytchev, B., Tamaki, T., Kaneda, K.: Improving the reliability of semantic segmentation of medical images by uncertainty modeling with Bayesian deep networks and curriculum learning. In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*, pp. 34–43. Springer (2021)
4. Linmans, J., van der Laak, J., Litjens, G.: Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks. In: *MIDL*. pp. 465–478 (2020)
5. Sahiner, B., Pezeshk, A., Hadjiiski, L.M., Wang, X., Drukker, K., Cha, K.H., Summers, R.M., Giger, M.L.: Deep learning in medical imaging and radiation therapy. *Medical physics* 46(1), e1–e36 (2019)
6. Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports* 7(1), 1–14 (2017)
7. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A.D.N., et al.: Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage* 195, 11–22 (2019)
8. Nair, T., Precup, D., Arnold, D.L., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Analysis* 59, 101557 (2020), <https://www.sciencedirect.com/science/article/pii/S1361841519300994>
9. Ozdemir, O., Woodward, B., Berlin, A.A.: Propagating uncertainty in multi-stage Bayesian convolutional neural networks with application to pulmonary nodule detection. *CoRR abs/1712.00497* (2017), <http://arxiv.org/abs/1712.00497>
10. Bhat, I., Kuijff, H.J., Cheplygina, V., Pluim, J.P.: Using uncertainty estimation to reduce false positives in liver lesion detection. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. pp. 663–667 (2021)
11. Mehrtash, A., Wells, W., Tempany, C., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging* PP, 1–1 (07 2020)
12. Hoebel, K., Andrearczyk, V., Beers, A., Patel, J., Chang, K., Depeursinge, A., Müller, H., Kalpathy-Cramer, J.: An exploration of uncertainty information for segmentation quality assessment. In: Išgum, I., Landman, B.A. (eds.) *Medical Imaging 2020: Image Processing*, vol. 11313, pp. 381–390. International Society for Optics and Photonics, SPIE (2020), <https://doi.org/10.1117/12.2548722>
13. Devries, T., Taylor, G.W.: Leveraging uncertainty estimates for predicting segmentation quality. *ArXiv abs/1807.00502* (2018)
14. Seeböck, P., Orlando, J., Schlegl, T., Waldstein, S., Bogunović, H., Riedl, S., Langs, G., Schmidt-Erfurth, U.: Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct. *IEEE Transactions on Medical Imaging* PP, 1–1 (05 2019)

15. Hiasa, Y., Otake, Y., Takao, M., Ogawa, T., Sugano, N., Sato, Y.: Automated muscle segmentation from clinical ct using Bayesian u-net for personalized musculoskeletal modeling. *IEEE Transactions on Medical Imaging* 39(4), 1030–1040 (2020)
16. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017)
17. Jungo, A., Reyes, M.: Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 48–56. Springer (2019)
18. Houlsby, N., Huszár, F., Ghahramani, Z., Lengyel, M.: Bayesian active learning for classification and preference learning. *arXiv preprint [arXiv:1112.5745](https://arxiv.org/abs/1112.5745)* (2011)
19. Smith, L., Gal, Y.: Understanding measures of uncertainty for adversarial example detection. *arXiv preprint [arXiv:1803.08533](https://arxiv.org/abs/1803.08533)* (2018)
20. Lu, S.L., Liao, H.C., Hsu, F.M., Liao, C.C., Lai, F., Xiao, F.: The intracranial tumor segmentation challenge: Contour tumors on brain mri for radiosurgery. *NeuroImage* 244, 118585 (2021)
21. van der Voort, S.R., Incekar, F., Wijnenga, M.M., Kapsas, G., Gahrman, R., Schouten, J.W., Dubbink, H.J., Vincent, A.J., van den Bent, M.J., French, P.J., et al.: The erasmus glioma database (egd): Structural mri scans, who 2016 subtypes, and segmentations of 774 patients with glioma. *Data in brief* 37, 107191 (2021)
22. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics* 38(2), 915–931 (2011)
23. Tsai, E.B., Simpson, S., Lungren, M.P., Hershman, M., Roshkovan, L., Colak, E., Erickson, B.J., Shih, G., Stein, A., Kalpathy-Cramer, J., et al.: The rsna international covid-19 open radiology database (ricord). *Radiology* 299(1), E204–E213 (2021)
24. Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.W., Han, X., Heng, P.A., Hesser, J., et al.: The liver tumor segmentation benchmark (lits). *arXiv preprint [arXiv:1901.04056](https://arxiv.org/abs/1901.04056)* (2019)
25. Pimkin, A., Samoylenko, A., Antipina, N., Ovechkina, A., Golanov, A., Dalechina, A., Belyaev, M.: Multidomain ct metal artifacts reduction using partial convolution based inpainting. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. pp. 1–6. IEEE (2020)
26. Saparov, T., Kurmukov, A., Shirokikh, B., Belyaev, M.: Zero-shot domain adaptation in ct segmentation by filtered back projection augmentation. In: *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections*, pp. 243–250. Springer (2021)
27. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* 18(2), 203–211 (2021)
28. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: *2016 fourth international conference on 3D vision (3DV)*. pp. 565–571. IEEE (2016)

# Towards Developing a Lightweight Neural Network for Liver CT Segmentation



Mohammed Yusuf Ansari, Snigdha Mohanty, Serah Jessy Mathew, Subhashree Mishra, Sudhansu Sekhar Singh, Julien Abinahed, Abdulla Al-Ansari, and Sarada Prasad Dakua

**Abstract** Image segmentation is crucial during the diagnosis and treatment planning of various liver diseases, especially hepatocellular carcinoma (hcc). We present a new neural network, Res-PAC-UNet, employing Pyramid Atrous Convolutions and a fixed-width residual UNet backbone resulting in low parameter count and of course, good liver CT segmentation. We use medical segmentation decathlon dataset to train the network. The resulting segmentation gives a Dice similarity coefficient of  $0.958 \pm 0.015$  with less than 0.5 million parameters with 1.2 million parameters.

**Keywords** Liver · Segmentation · Neural networks

## 1 Introduction

Outlining the human organs on medical images helps in proper planning and prevent the clinicians from damaging the surrounding tissues. Furthermore, the segmented images can have other applications, such as, in image fusion of ultrasound (US), magnetic resonance image (MRI), computed tomography (CT), etc. to enhance visualization. The fused images can be used in image guided surgeries or interventions. There have been mainly two types of segmentation, automatic and semi-automatic. The automatic ones face various challenges due to the nature of the method itself and that of the complexity of intensity distribution caused by the cancer [1–3]. A few challenges are: (1) the intensity distribution between liver and surrounding tissues is such that sometimes it would be difficult for a non-clinician to discriminate, (2) use of contrast enhancement sometimes results in increased noise level and artifacts on the CT scans, (3) voxel spacing and axial resolution on CT scans cause loss of necessary volumetric information for CT segmentation [4], (4) the liver tumors sometimes

---

M. Y. Ansari · S. J. Mathew · J. Abinahed · A. Al-Ansari · S. P. Dakua (✉)  
Department of Surgery, Hamad General Hospital, Hamad Medical Corporation, Doha, Qatar  
e-mail: [sdakua@hamad.qa](mailto:sdakua@hamad.qa)

S. Mohanty · S. Mishra · S. S. Singh  
School of Electronics, KIIT Deemed to be University, Bhubaneswar, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
R. Su et al. (eds.), *Medical Imaging and Computer-Aided Diagnosis*, Lecture Notes  
in Electrical Engineering 810, [https://doi.org/10.1007/978-981-16-6775-6\\_3](https://doi.org/10.1007/978-981-16-6775-6_3)



have complex features, such as varying intensity, varying shape and size making the segmentation task daunting, and (5) additionally, the partial volume effect can't be neglected adding further ambiguity.

The literature in image segmentation is quite rich; still, new methods keep coming to overcome the challenges posed by the input data and segmentation methods. Earlier, these techniques include: region growing [5–7], model-based [8], clustering [9], graph cut [10, 11], etc. However, the conventional methods have failed to perform as per the expectation level of the clinicians, especially with respect to accuracy, robustness, and automation, thus, neural network has presently taken the limelight [12].

The convolutional kernels in the neural network extract the relevant features in the input data (image) reducing the user dependency and increasing the accuracy. Ronneberger et al. [13] has revolutionized by proposing a network that has been so popular that the modifications keep coming in every now and then. This is based on an encoder-decoder concept, where the encoder learns to generate a dense feature representation from the input data and the decoder creates the segmentation mask. Repeated pooling causes a loss in spatial information; thus, the skip connection has been introduced to minimize this loss.

Recently, a new network has been proposed, Thin-UNet [14], that achieves image segmentation with less parameter count. In this paper, we especially focus on reducing the model size, parameter count, and model usability keeping appropriate acceptable segmentation accuracy.

## 2 Proposed Methodology

### 2.1 Network Architecture

Although UNet has been quite popular, there are some limitations with respect to skip connections and large parameter counts [15]. Duplications of low resolution feature maps are resulted from the encoder (E) and propagated to decoder (D) leading to smoothing the object boundaries. Thus, to overcome this problem, a novel network, Res-PAC-UNet (Fig. 1), is proposed, where constant feature width (W) of tuned backbone in addition to the residual (R) blocks are used minimizing the memory footprint and parameter count. This also improves the gradient and information flow. We employ strided convolutions; the convolutional blocks are replaced by the R blocks; this downscales the input features. We also aim to leverage multi-scale volumetric features from the low-resolution feature maps of the encoder; Pyramid Atrous Convolution Module (PAC) modules is deployed over the skip connections to generate these. PAC modules are not placed at the top of the skip connections to avoid large memory that is required for high resolution feature maps. It is shown in Fig. 2 on how the R block is being used in the tuned backbone. The feature map is downscaled by half due to the initial convolutions in E residual blocks with a stride of 2 ( $s_0 = 2$ ). On the other hand, upscaling of the feature map occurs by the decoder

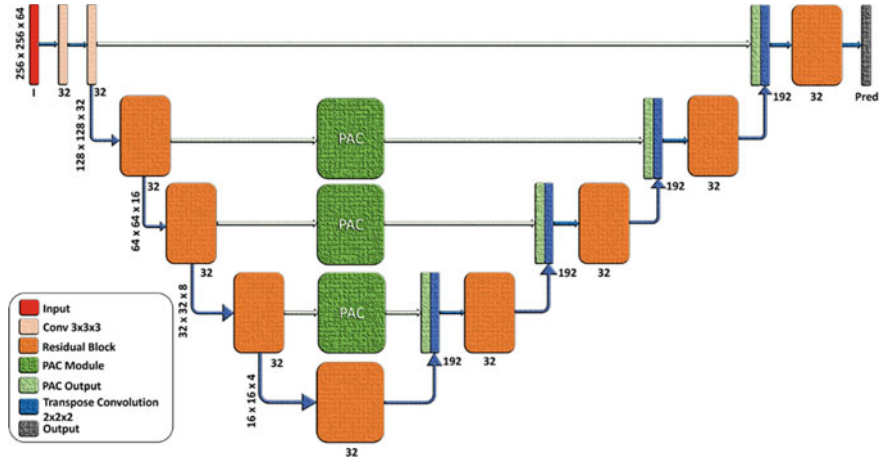


Fig. 1 The Res32-PAC-UNet network for segmenting liver CT

after transposing convolutions and regular convolutions, where a 1 stride does the job in the R blocks. The expression for regular convolutional operation and R blocks may be provided as:

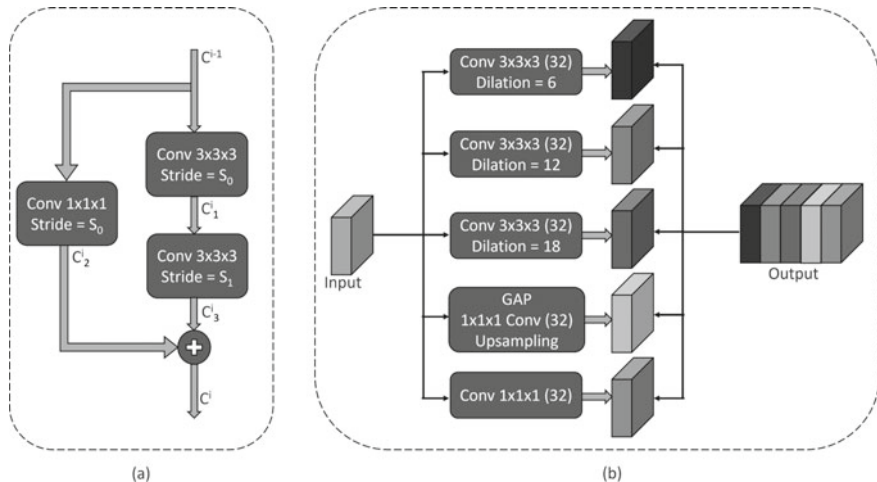
$$\text{Conv}_{-u \times u \times u}(xi, s, W; \theta) = f(w^j \otimes_s xi + b^j), \forall 1 \leq j \leq W, w^j \in \theta, b^j \in \theta, \quad (1)$$

where,  $xi, s, W$ , and  $u$  represent feature map, convolution stride, number of kernels, kernels dimension, respectively; furthermore, the information on kernels weights and biases are contained by  $\theta$ . Additionally, the activation function,  $f(\cdot)$ , is applied to the convolution result,  $\otimes_s$  is the result of strided convolution operation.  $b^j$  and  $w^j$  represent the  $j$ th kernel bias and the  $j$ th kernel weight, respectively. Thus, the R block can be expressed as:

$$\begin{aligned} o_1^i &= \text{Conv}_{-u \times u \times u}(c^{i-1}, s_0, W; \theta_1^i), \\ o_2^i &= \text{Conv}_{-u \times u \times u}(c^{i-1}, s_0, W; \theta_2^i), \\ o_3^i &= \text{Conv}_{-u \times u \times u}(c_1^i, s_1, W; \theta_3^i), \\ o^i &= o_2^i \oplus o_3^i, \end{aligned} \quad (2)$$

where,  $o^{i-1}$  and  $o^i$  are the R block input and output, respectively.  $o_1^i, o_2^i, o_3^i$  are the 3 convolutional operations, whereas,  $\oplus$  represents addition operation element-wise.

**Modified Surface Loss** There has been some popular loss functions, such as focal loss, surface loss, binary cross-entropy (BCE) and others [18]. However, the distance metrics do play crucial role to quantify the boundary errors, thus, Kervadec et al. [16] present a boundary loss function that describes a graph-based optimization to estimate



**Fig. 2** **a** Residual block used to improve the gradient flow and information. **b** PAC building block to capture the volumetric features of multi-scale nature at the encoder side

the flow of the gradient for curve evolution. Subsequently, the regional softmax probabilities of the pixels ( $\Omega$ ) are used to calculate the boundary loss in the ground truth level-set function ( $\phi_G$ ) and predicted segmentation mask ( $M_\theta$ ).

$$BL(\Omega) = \int_{\Omega} \phi_G(p) M_\theta(p) dp. \quad (3)$$

Kervadec et al. [16] combine the region-based loss (surface loss) with boundary loss providing improvement in accuracy of 8%. Thus, we have modified the above proposed loss function; we replace with a combo loss that is the sum of Dice loss and focal loss. The objective is to emphasize the distribution with regards to area and class of the regions of interests (ROI). This would certainly improve corresponding metrics of class accuracy. In addition, a weight shifting strategy has been proposed that shifts the weight from 0.01 to 0.75, and 0.99 to 0.25, of boundary loss and combo loss, respectively. This strategy helps in achieving a decent portion of the net weight after the training.

### 3 Setup for the Experiment

#### 3.1 Data

In medical segmentation decathlon [17], liver CT scans were used to train the models. The database accommodates 201 CT scans that are contrast-enhanced with 131 scans

in the train set and 70 Scans in the test set. The scans' spatial dimension was  $512 \times 512$ , and number of slices was having a range (of 50, 1100). The scans are from patients, who have hcc or liver metastases. Performing training (101 scans) and validation (31 scans) on every original set was done on the CT scans to overcome the challenges in the test set.

Nifti loader was used to read the file during the pre-processing stage and the scans that were in the range  $[-500, 500]$  HU were captured. The image intensities were then recomputed to  $[-1, 1]$  by min-max normalization. To reduce VRAM consumption, spatial measurements of input scans were resized to  $256 \times 256$ , and 64 slices from each scan's liver territory were resampled. The consumption of VRAM is a major issue, when designing networks for 3D CT. The tumor label was then replaced with liver label for training liver CT segmentation networks. *Volumentations* package was applied to the refined CT scans to minimize overfilling.

### 3.2 Implementation Details

CT scans were stored in RAM before training to reduce input/output (I/O) as well as computational costs. Neural networks in Keras<sup>1</sup> was also built using Tensorflow dataset generator and prefetching, which makes certain the neural networks were reliably supplied enhancing the scans together with ground truth. To establish model convergence, the networks were trained for 150 epochs (Fig. 3). Three different loss functions and modified surface loss function were used for training Res32-PAC-UNet and other models, respectively. For network parameter updates, batch size of 1 and the Adam optimizer (learning rate = 0.0001) were used. The Keras callbacks were used to save the model weights that produced an elevated Dice coefficient (DC) on the test set and they were then applied for evaluating the model. We have used a workstation (HP Z8), whose specifications are as follows: Intel®Xeon(R) Silver, 4216 CPU, 64 cores, 2.10GHz base clock, and 128 GB of system memory.

## 4 Results

Table 1 displays the Res32-PAC-UNet model's performance summary for three distinct loss functions. It depicts that using binary cross-entropy as well as focal loss results in less reliable segmentation. This can be mitigated by focusing on the significance of area/volume overlap information in training networks for segmentation tasks, which increases segmentation reliability, and by region overlapping along with class distribution in modified surface loss.

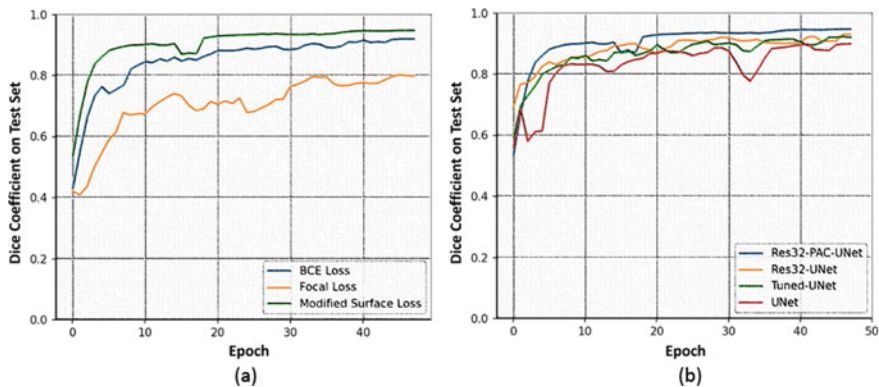
The Res32-PAC-UNet model's 3-moving average DC, which is primed for the initial 100 epochs utilizing disparate loss functions is shown in Fig. 3. The sudden

---

<sup>1</sup> F. Chollet et al., <https://github.com/fchollet/keras>, 2015.

**Table 1** Quantitative performance of Res32-PAC-UNet with respect to different loss functions, the bold figures indicate the significance of modified surface loss function

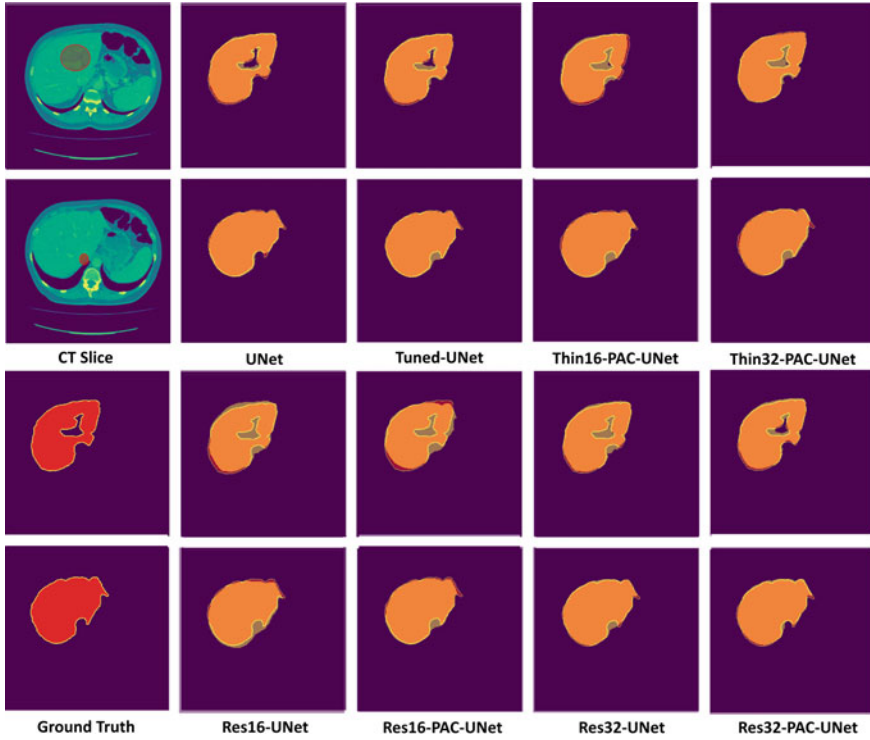
Loss function	DC	IoU	Sensitivity	Specificity	SVD	VOE
Focal loss	0.898 (0.024)	0.815 (0.038)	0.95 <b>(0.023)</b>	<b>0.998</b> (0.002)	0.102 (0.024)	0.185 (0.038)
Binary cross entropy	0.949 (0.016)	0.903 (0.028)	0.965 (0.028)	0.997 <b>(0.001)</b>	0.051 (0.016)	0.097 (0.028)
Modified surface loss	<b>0.958</b> <b>(0.015)</b>	<b>0.92 (0.026)</b>	<b>0.96</b> (0.026)	0.997 <b>(0.001)</b>	<b>0.042</b> <b>(0.015)</b>	<b>0.08 (0.026)</b>



**Fig. 3** On the test set, DC evolution in first 50 epochs of training: **a** the network trained with the loss functions. **b** The network trained with the proposed loss function

changes in the DC curve produced due to stochastic network weight updates were smoothed out by applying moving averages. By reaching an 80% DC in the first initial epochs, modified surface loss offers high segmentation accuracy and speedy convergence in earlier epochs. Res32-PAC-UNet can achieve excellent segmentation accuracy due to the modified surface loss, which also hastens convergence among the tested loss functions (Fig. 4).

An important feature of Res-PAC-UNet model is that it was designed for maximizing segmentation accuracy while reducing parameter count and disk utilization. The Tuned-UNet overcomes UNet model's 270 MB model weight by reducing parameter counts and storage space by up to 4 $\times$ . Parameter reduction accelerates segmentation completion, increasing DC from 91.9 to 95.5% when compared to UNet. The Thin16-PAC-UNet, as well as Thin32-PAC-UNet models, advance towards Tuned-UNet model's segmentation conduct; the figures reach as much as 12 $\times$ , 4.6 $\times$  fewer parameters and storage requirements, respectively. Because of the PAC module in thin fixed-width architectures, the segmentation performance of Thin-PAC-UNet was found as better than Tuned-UNets. By confining the model size to 15.1 MB and the parameters to 1.2 million, the Res32-PAC-UNet out-stands other versions



**Fig. 4** A qualitative comparison of the proposed neural network. The artifacts are marked by red bounding oval marks. Overlapping of the predicted segmentation masks (yellow) on the ground truth (red)

in the empiric study. The newly proposed Res-UNet++ [19] architecture has nearly identical performance to Res32-PAC-UNet, but has almost  $10\times$  more parameters. On the other hand, if segmentation accuracy was the goal, the Res32-PAC-UNet is agreed upon more than UNet and Tuned-UNet models, as it attains best accuracy with  $18\times$ ,  $4.6\times$  fewer parameters.

## 5 Conclusion

It was suggested to employ an original Res-PAC-UNet architecture that combines PAC modules with a customized fixed-width R backbone to achieve good segmentation performance with minimal weights. The PAC modules located over the skip-connection are helped in extracting pertinent multi-scale volumetric features by the R backbone, which limits the exponential growth rate of the parameters while enhancing information and gradient flow. We have modified the surface-based loss func-

tion and trained the network in order to enhance the performance of the segmentation. Res32-PAC-UNet has proved to have maximized the segmentation performance.

**Acknowledgements** This publication was made possible by NPRP- 11S-1219-170106 from the Qatar National Research Fund (a member of Qatar Foundation). The findings herein reflect the work, and are solely the responsibility of the authors. This research was also co-funded and supported by the Medical Research Center, Hamad Medical Corporation, Doha, Qatar.

## References

1. Dakua, S., & Sahambi, J. S., Weighting Function in Random Walk Based Left Ventricle Segmentation, Proc. of 18th IEEE International Conference on Image Processing, Brussels (Belgium), 2133–2136, 2011.
2. Rai, P., Abin角度, J., Dakua, S., & Balakrishnan, S., Feasibility and efficacy of fusion imaging systems for immediate post ablation assessment of liver neoplasms: Protocol for a rapid systematic review, International Journal of Surgery Protocols, IJS Press, 25, 1, 209–215, 2021.
3. Sarada, D., & Nayak, A., A Review on Treatments of Hepatocellular Carcinoma - Role of Radio Wave Ablation and Possible Improvements, Egyptian Liver Journal, Springer, 12, 30, 1–10, 2022.
4. Akhtar, Y., Dakua, S., Abdalla, A., Aboumarzouk, O., Ansari, M. Y., Abin角度, J., Elakkad, M. S. M., Al-Ansari, A., Risk Assessment of Computer-aided Diagnostic Software for Hepatic Resection, IEEE Transactions on Radiation and Plasma Medical Sciences, 2021, <https://doi.org/10.1109/TRPMS.2021.3071148>.
5. Singh, A., Romeo, A., Scott, K., Wagener, S., Leibrock, L., Laux, P., Luch, A., Kerkar, P., Balakrishnan, S., Dakua, S., & Park, B., Emerging technologies for in vitro inhalation toxicology, Advanced Healthcare Materials, Wiley, 10, e2100633, 2021.
6. Dakua, S., & Sahambi, J., Modified Active contour Model and Random Walk Approach for Left Ventricular Cardiac MR Image Segmentation, International Journal for Numerical Methods in Biomedical Engineering, Wiley, 27, 1350–1361, 2011.
7. Singh, A., Laux, P., Luch, A., Balkrishnan, S., & Dakua, S., Bottom-UP assembly of nanorobots: extending synthetic biology to complex material design, Frontiers in Nanoscience and Nanotechnology, 5, 1–2, 2019.
8. Singh, A., Maharjan, R., Kromer, C., Laux, P., Luch, A., Vats, T., Chandrasekar, V., Dakua, S., & Park, B., Advances in smoking related in-vitro inhalation toxicology: a perspective case of challenges and opportunities from progresses in lung-on-chip technologies, ACS Chemical Research in Toxicology, 34, pp. 1984–222, 2021.
9. Dakua, S., & Sahambi, J. S., LV Contour Extraction from Cardiac MR Images Using Random Walk Approach, In IEEE International Advance Computing Conference, Patiala, India, 228–233, 2009.
10. Dakua, S., Abin角度, J., & Al-Ansari, A., A PCA based Approach for Brain Aneurysm Segmentation, Journal of Multi Dimensional Systems and Signal Processing, Springer, 257–277, 2018.
11. Dakua, S., Abin角度, J., & Al-Ansari, A., Pathological Liver Segmentation Using Stochastic Resonance and Cellular Automata, Journal of Visual Communication and Image Representation, ScienceDirect, 34, 89–102, 2016.
12. Ansari, M. Y., Abdalla, A., Ansari, M. Y., Ansari, M. I., Malluhi, B., Mohanty, S., Mishra, S., Singh, S. S., Abin角度, J., Al-Ansari, A., Balakrishnan, S., & Dakua, S. P., Practical utility of liver segmentation methods in clinical surgeries and interventions, BMC Medical Imaging, 22, 97, 1–17, 2022.

13. Ronneberger, O., Fischer, P., & Brox, T., U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham, 2015.
14. Vaze, S., Xie, W., & Namburete, A., Low-Memory CNNs Enabling Real-Time Ultrasound Segmentation Towards Mobile Deployment, *journal of biomedical and health informatics, IEEE*, 24, 4, 1059–1069, 2020.
15. Ansari, M., Yang, Y., Balakrishnan, S., Abinahed, J., Al-Ansari, A., Warfa, M., Almokdad, O., Barah, A., Omer, A., Singh, A., Meher, P., Bhadra, J., Halabi, O., Azampour, M., Navab, N., Wendler, T., & Dakua, S., A Lightweight Neural Network with Multiscale Feature Enhancement for Liver CT Segmentation, *Scientific Reports, Nature*, 12, 14153, 1–12, 2022.
16. Kervadec, H., Bouchtiba, J., Desrosiers, C., Dolz, E., & Ayed, I., Boundary loss for highly unbalanced segmentation. In *International conference on medical imaging with deep learning*, 285–296 (PMLR, 2019).
17. Antonelli, M., Reinke, A., & Bakas, S., *et al.*, The Medical Segmentation Decathlon, *Nat Communication*, 13, 4128, 2022. <https://doi.org/10.1038/s41467-022-30695-9>
18. Ramos, D., Javier, F., Alicia, L., & Joaquin, G., Deconstructing Cross-Entropy for Probabilistic Binary Classifiers, *Entropy* 20, 3, 208, 2018.
19. Jha, D., Smedsrud, P., Riegler, M., Johansen, D., Lange, T., Halvorsen, P., & Johansen H., Resunet++: An advanced architecture for medical image segmentation, In *2019 IEEE International Symposium on Multimedia (ISM)*, 225–230, 2019.



# NuRISC: Nuclei Radial Instance Segmentation and Classification



Esha Sadia Nasir and Muhammad Moazam Fraz

**Abstract** Accurate segmentation and classification of nuclei instances is one of the most challenging tasks due to wide occurrence of overlapping, cluttered nuclei having blurred boundaries. Existing methods particularly focus on region proposal techniques and feature encoding frameworks, however often fails to precisely identify instances. In this paper we propose a simple yet effective model that precisely recognize instance boundaries as well as caters exhaustive class imbalance problems, thus yielding accurate class information for each nuclei. We have utilized nuclei pixel positional information i.e. its distance from contours for accurate shape estimation along with an object probability score for filtering true nuclei pixels from background. The network comprises of a light weight multi head U-Net architecture having separate instance probability, shape radial estimator and classification heads. A compound classification loss function is used that minimizes loss by assigning weighted loss to each class according to type occurrence frequency thus mitigating major class imbalance issues existing in most of publicly available nuclei datasets.

**Keywords** Whole slide imaging · Nuclei · Segmentation · Computational pathology · Deep learning · Classification · Detection

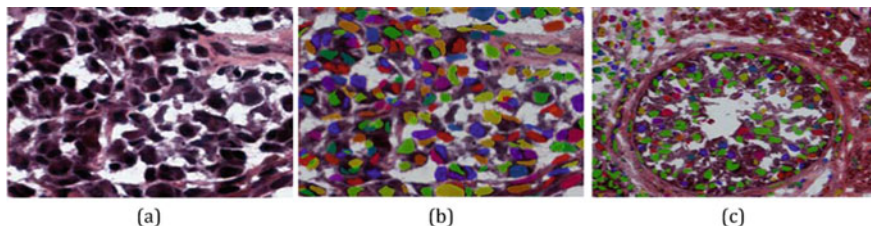
## 1 Introduction

Accurate segmentation and classification of nuclei is considered as a preliminary step towards an intricate whole slide image analysis leading to circumstantial histology images research. For instance, nuclei counts on digital pathology images have noteworthy diagnostic importance in various cancerous stage particularly including cancer grading, phenotyping, patients survival prediction, automatic nuclear pleomorphism scoring and mitosis detection. All of these deliberately relies on nuclei instance appearances and structural variations [1]. Nuclei presence, size, shape, staining and morphological characteristics are important indicators in estimation of

---

E. S. Nasir (✉) · M. M. Fraz

National University of Sciences and Technology (NUST), Islamabad, Pakistan  
e-mail: [enasir.mscs20seecs@seecs.edu.pk](mailto:enasir.mscs20seecs@seecs.edu.pk)



**Fig. 1** Adrenal gland WSI from CryoNuSeg dataset along with labelled ground truth indicating high number of occluded nuclei **a** image, **b** ground truth and **c** occlusion

diseases severity. However, manually segmenting such structures is tedious as well as error prone due to extreme inter as well as intra observer variability. Figure 1 shows high number of occluded and overlapping nuclei from adrenal gland CryoNuSeg Dataset. These occlusion later on hinders model training and yields poor performance [2]. Contrary to this automated methods that reports high performance on a particular WSIs data yields poor results on distinct datasets due to disparity in organic cells with respect to different organ tissues as well as variation in acquisition parameters including color inconsistency due to staining variations and occluded nuclei boundaries [3]. Similarly, malignant cells growth rate is extremely high and it's density in malignant cells is also reported much higher compared to normal cells. Squeezing these two often times yield large number of clumped nuclei instances [4].

## 2 Related Work

Nuclei segmentation and classification is an elemental task in computer aided disease diagnosis and tumor micro environment analysis [5, 6]. Traditional approaches used for segmentation of nuclei comprises of thresholding, watershed [7] segmentation [8], level-sets [9], morphological operations [10], active contour models [11] and snake energy optimizations [12]. A notable shortcoming of all these handcrafted techniques is inadequacy to fully detect nuclei due to it's dependency on low-level features lacking significant structural details thus leading to degraded segmentation results [13, 14]. In past few years, convolutional neural networks based deep learning techniques have surpassed traditional methods in nuclei instance segmentation [15]. In 2017 Kumar et al. [16] proposed a convolution neural network based on pixels classification, against every image pixel a probability score is computed yielding 3 class output information for nuclei boundary, interior and exterior probabilities. In 2018 [17] proposed a star convex polygons based cell localization network considering better shape representation results of convex polygons compared to usual bounding box based detection and thus do not need shape refinement. For this, they trained a convolutional neural network for predicting every pixel within that polygon cell instance at that position. Similarly, Graham et al. [18] proposed distance based

nuclei identification and classification technique where nuclei instances estimation is done using pixel to nuclei centroid distance maps in horizontal as well as vertical directions. A joint attention model based on Neural Architecture Spatial and channel weighting effect is proposed by Liu et al. [19] using NAS search strategy for attention module automation with the addition of multiple attention module architectures searching within same network.

Above mentioned approaches though yield state of the art results however for final segmentation instance primarily uses sophisticated post-processing modules including watershed [20], conditional random fields (CRFs), morphological erosions or dilations and clustering [21]. Recently, shape aware nuclei identification techniques have been proposed, where a polygon is used for representing each individual instance and is calculating via nuclei center and boundary pixels prediction.

### 3 Methodology

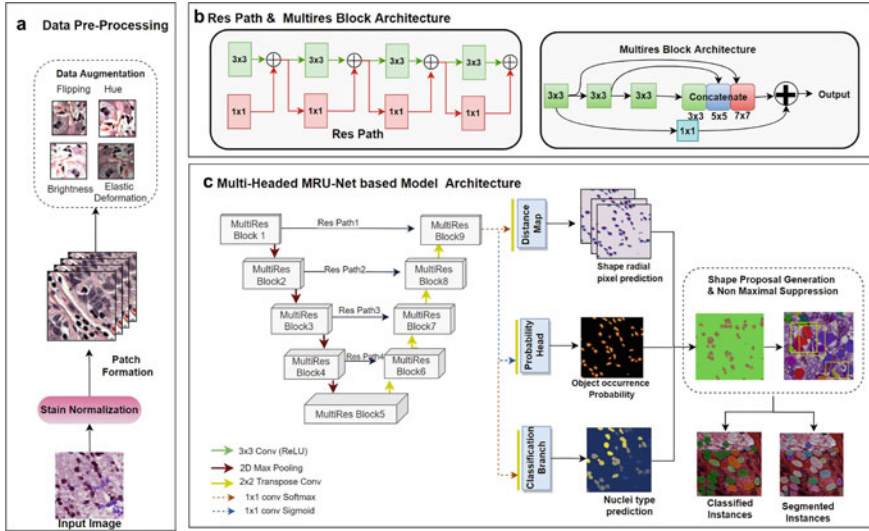
Nuclei segmentation and classification is one of the basic step yielding rich information towards further cancer research including cancer grade estimation, phenotyping, quantification and survival prediction. For exploiting these details, in this paper we proposed a single stage probabilistic model for multi class nuclei instance segmentation and type classification. The main aim of the architecture is identification of nuclei instances using center of mass and contour features information. For pixel at any point it calculates estimated inter space towards the nuclei edges using arc gradients metrics. The block diagram of the framework is shown in Fig. 2.

#### 3.1 Deep Regression Network

This model is an extension of the state of the art MRU-Net [22] encoder decoder based segmentation network that regresses not just instance locations, confidence scores and class probabilities for each instance, along with detailed shape encoding.

#### 3.2 Instance Representation

We require a compact and interpretable object embedding representation embodying higher understanding of each instance shape and overlapping patterns for the nuclei shapes prediction. For this direction maps within each instance yields morphometric information via decodable shape representations to radial vectors and object probability and finally learned shape encodings.



**Fig. 2** **a** Shows data preparation and preprocessing using different augmentation techniques. **b** Shows MRU-Net building block including ResPath and Multires blocks architecture. **c** Show NuRiSC architecture where preprocessed patches are given to Multiheaded MRU-Net as input and it outputs distance maps, probability head and classification outputs followed by Shape proposal generation and Non maximal suppression

### 3.3 Radial Distance Maps

Radial distance maps represents each instance as a line segment from a hypothetical centre spot within the object and directs pixel distributed over nuclei contours. Model finds contour points via finding pixels where radial direction cast away from the central point intersecting the boundary at angles in range from  $0$  to  $2\pi$ . For finding optimal shape we construct multiple radial directions for every nuclei instance and finally selects one on the basis of maximum IoU threshold.

### 3.4 Pre-processing

For enhancing model accuracy with fewer learning parameters we have applied pre-processing for data preparation. During pre-processing stage, Structure preserving color normalization is used for mitigating redundant variations including image contrast and differences in reagent concentrations during scanning. Similarly, for better training data augmentation including rotations, horizontal and vertical flips and intensity variations are applied.

### 3.5 Model Architecture

Initially, high and low resolution feature maps are generated via a backbone convolutional neural network with 3 outputs prediction heads, each consisting of single convolution layer yielding three specific details per pixel including (1) instance shape direction map, (2) instance probability score and (3) class information respectively. The CNN backbone is comprised of 5 levels having 16, 32, 64, 128 and 256 channels. Similarly each encoding, decoding block comprises of two  $3 \times 3$  convolution blocks along with batch Normalization, ReLU activation function and a  $2 \times 2$  max-pooling or upsampling branch. The generated feature map has 256 channels and ReLU activations. In probability estimation head single channel output with sigmoid activation is used for separating object from background. Similarly, in the shape estimator branch  $n$  output channels are produced for  $n$  vertex polygonal nuclei along with ReLU activation. In classification head  $m$  channel output with soft max activation is used where  $m$  indicates number of nuclei classes. In next stage pixels having probability score greater than thresholds are selected for nuclei shape formation. Similarly for instances classification an additional classifier head is used in backbone CNN along with other two heads including object existence probability head and shape estimator head. Similar, to object probability head, classifier head yields class probability of pixel which is finally aggregated for all pixels thus representing finalized class instance. Due to pixels surrounding multiple objects and voting for  $n$  instances simultaneously, an IoU based non-maximal suppression is used for removing redundant instances while keeping most matched one thus mitigating multiple similar instance formations via eliminating false positive candidates. In this stage out of several objects the one above a specific threshold is selected i.e. candidate having best normalized intersection over union overlap threshold.

### 3.6 Loss Functions

**Classification Loss:** For alleviating huge class imbalance issue in majority of our training datasets, we used a joint categorical Tversky loss metric that assign larger weights to less frequent class pixels similarly, minimal weight to class having higher occurrence. Loss is computed separately for initially for categorical cross entropy and Tverksy and finally mean of generalized combined Dice and Cross-Entropy Loss also regarded as unified focal loss [23] is returned thus minimizing imbalance effects.

$$L_{FTL} = \left( 1 - \frac{TP}{TP + \alpha FN + \beta FP} \right)^{\gamma} \quad (1)$$

$$L_{WCE} = -wt_i \log(P_i) \quad (2)$$

$$L_{cls} = L_{WCE} + L_{FTL} \quad (3)$$

**Regression Loss:** For estimation of each instance radii we have applied mean absolute error(mae)loss functions, Similarly for object existence probability computation we have used binary cross entropy loss function.

$$L_{\text{dist}} = \frac{1}{N} \sum_{k=0}^{N-1} r_{ij}^k - r_{ij}^k \quad (4)$$

$$L_{\text{prob}} = \frac{1}{N} \sum_{i=1}^N -(y_i * \log(p_i) + (1 - y_i) * \log(1 - p_i)) \quad (5)$$

### 3.7 Post-processing

In post processing, like other detection based methods, from multiple radii based generated proposals, NuRISC removes redundant ones during inference via applying IoU based non maximal suppression (NMS). It basically keeps proposals with higher score while suppresses the ones with lower score and IoUs exceeding the specified thresholds.

## 4 Experiments and Results

### 4.1 Implementation Details

For all experiments, we have used a work station equipped with an Intel Core i9 CPU, 32 GB RAM and GeForce V100 GPU. All experiments are done in Keras framework having Tensorflow backend. For all applications, NuRISC is trained for 300 epochs. Adam optimizer with learning rate of  $3 \times 10^{-4}$  and weight decay of half after every 40 epoch was used during model training. Batch size of 4 is used for all datasets. We have used following augmentation operations including: random flipping (horizontal and vertical), elastic deformation, hue and brightness adjustment.

### 4.2 Evaluation Metrics

For validation study and testing, we have use metrics that are reported in the literature for nuclei instance segmentation and classification. Following five measures have been used for comparative analysis of segmentation performance evaluation for different models. Including Accuracy, Precision, Recall, F1-Score and Panoptic

**Table 1** Publicly available datasets used for training

S. no	Dataset	# Nuclei	Mag	Organs	Source
1	CoNSeP [18]	24,319	40×	1	UHCW <sup>a</sup>
2	PanNuke [24]	205,343	40×	19	TCGA <sup>a</sup>
3	CryoNuSeg [26]	7596	40×	10	TCGA <sup>b</sup>
4	CPM-17 [27]	7570	40×	1	TCGA <sup>a</sup>
5	CPM-15 [27]	2906	20×	1	TCGA <sup>a</sup>
6	TNBC [26]	4022	40×	1	TCGA <sup>a</sup>
7	Kumar [25]	21,623	20×	7	TCGA <sup>a</sup>

<sup>a</sup>The Cancer Genome Atlas

<sup>b</sup>University Hospital Coventry and Warwickshire

Quality metric. Accuracy basically indicates the overall classification accuracy. Similarly, Precision and Recall is True positive rate of identified instances while F1 score is the weighted average of both Precision and Recall. For final performance evaluation we used Panoptic Quality metric PQ which basically comprises of sum of F1 score i.e. detection quality DQ (H.M of Precision and Recall) and Segmentation Quality SQ i.e. average IoU for all accurately matched instances. panoptic quality PQ is defined as the product of detection quality DQ (F1 score, i.e. the harmonic mean of precision and recall) and segmentation quality SQ (average intersection over union of all correct matches).

### 4.3 Datasets

In this paper, we train and evaluate our proposed architecture on following publicly available nuclei instance segmentation datasets including PanNuke [24], CoNSeP <sup>1</sup>, Kumar [25], CryoNuSeg [26], TNBC [17], CPM15 and CPM-17 [27] datasets. Table 1 shows organs, nuclei counts, magnification of datasets used in this paper. In Fig. 5 represent tissue wise datasets distribution in major datasets.

### 4.4 Baseline Methods

We have evaluated performance for the following state of the art models.

- **Mask-RCNN:** It is one of the most frequently used 2-stage instance segmentation network proposed by He et al. [28] that generates region proposal for each target object, applies non-maximum suppression for filtering and eventually yielding masks for each object.

---

<sup>1</sup> University Hospital Coventry & Warwickshire.

- **U-Net:** This network is based on an encoder-decoder based approach with lateral skip-connections developed primarily for medical image analysis. In addition to initial framework proposed by Ronneberger et al. [22] after each conv layer we utilized batch normalization and finally classifying pixels into separate categories.
- **Hover-Net:** It is the first simultaneous segmentation and classification architecture that uses 3 separate branches for semantic segmentation, clustered nuclei separation and nuclei classification. For overlapping nuclei separation they have utilized horizontal and vertical distances w.r.t center of mass. We have used official repo of Graham et al. [18] that is available on GitHub.

## 4.5 Experimental Results

The results section is divided in two main parts including instance segmentation qualitative and quantitative results in first part while classification result in second part.

**Instance Segmentation Results** Table 2 shows the quantitative results comparison of already existing state of the art networks and our proposed technique in terms of Panoptic quality, precision, recall and dice similarity for all images from mentioned datasets. Figure 3 shows predictions visualization and comparison on sample nuclei patches with state of the art methods.

**Classification Results:** As shown in Table 3, the proposed model NuRiSC achieves state of the art results not only on in instance segmentation but classification as well. In particular, our method outperformed the best method till now i.e. HoVer-Net in terms of bPQ and mPQ across all datasets (Fig. 4).

Figure 5 represents huge class imbalance issue in PanNuke and CoNSEP datasets causing poor results for minority classes. We have catered this issue via using training class weights on the basis of class frequency with respect to total number of instances of each class and compound classification loss yielding better results compared to previously proposed classification architectures.

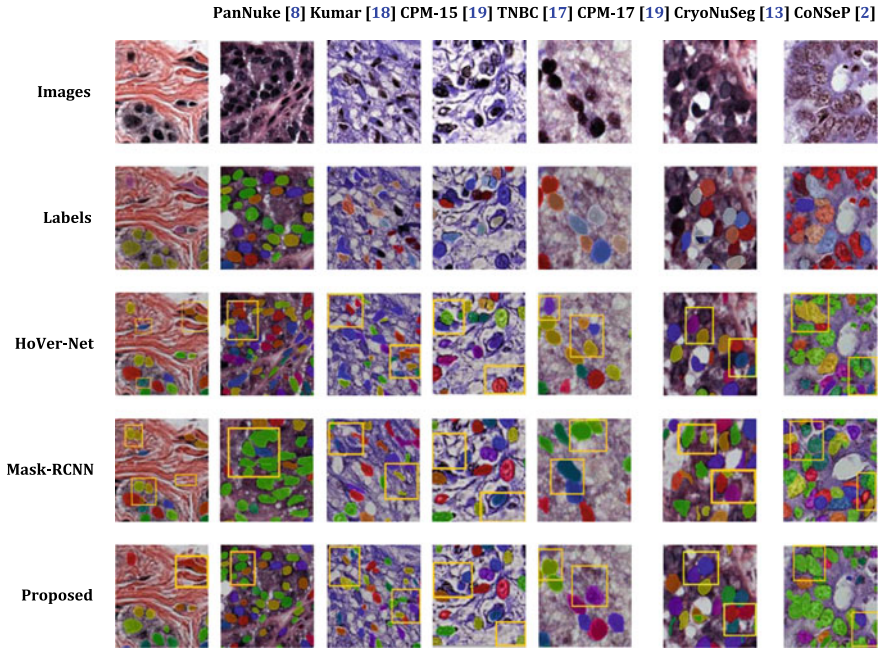


**Table 2** Instance segmentation Panoptic Quality (PQ), Precision (Pr), Recall(Re) and Dice Similarity Score (DSc) results comparison of proposed architecture with baseline networks

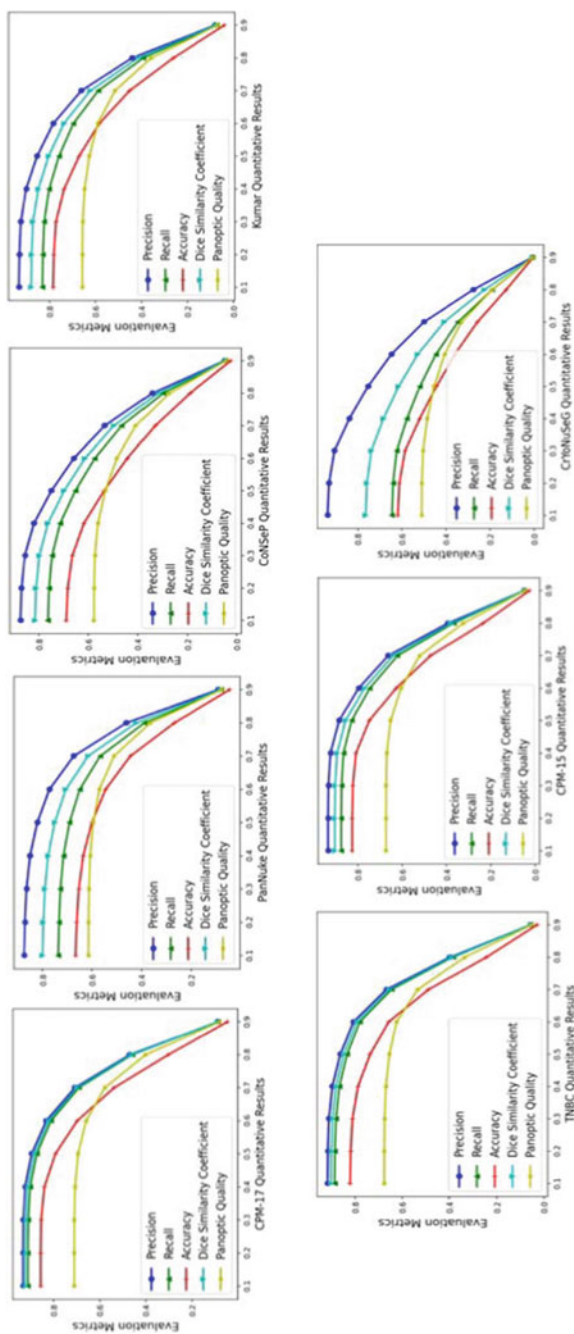
Datasets	Methods	PQ	Pr	Re	DSc
CoNSeP [18]	Mask R-CNN [28]	0.46	–	–	0.74
	U-Net [22]	0.33		–	0.72
	Hover-Net [18]	0.547			<b>0.85</b>
	Proposed	<b>0.54</b>	0.80	0.70	0.74
PanNuke [24]	Hover-Net [18]	0.46	0.82	0.79	<b>0.80</b>
	TsFD-Net	0.4456	–	–	–
	MaskR-CNN	0.3688	0.76	0.68	0.72
	Proposed	<b>0.61</b>	0.83	0.72	0.77
CryoNuSeg	Mask R-CNN [28]	0.39	0.63	0.54	0.63
	U-Net	0.38	0.62	0.51	0.64
	Proposed	0.53	<b>0.77</b>	<b>0.67</b>	0.72
CPM-17 [24]	Mask R-CNN [28]	0.67	–	–	0.85
	HoVer-Net	0.69	–	–	0.86
	Proposed	<b>0.70</b>	0.89	0.87	<b>0.89</b>
Kumar [25]	Mask R-CNN [28]	0.509	–	–	0.76
	U-Net [18]	0.58	–		0.478
	HoVer-Net	0.597	–	–	<b>0.82</b>
	Proposed	<b>0.63</b>	0.85	0.76	0.80
TNBC [17]	Mask R-CNN [28]	0.443	–	–	0.705
	U-Net [18]	0.442	–	–	0.681
	HoVer-Net	0.578	–	–	0.749
	Proposed	<b>0.65</b>	0.86	0.83	<b>0.85</b>
CPM-15 [27]	Mask R-CNN [28]	0.549	–	–	0.764
	U-Net [18]	0.446	–		0.720
	HoVer-Net	0.606	–	–	0.801
	Proposed	<b>0.65</b>	0.88	0.83	<b>0.86</b>

The bold values indicates the best performance result in comparison to all model results

Figure 6 shows classification results for PanNuke and CoNSeP datasets having example input images from the dataset, Hover-Net, Mask-RCNN and proposed predictions (Left to Right). Each nuclei instance color reflects its specific class labelled at the bottom.



**Fig. 3** Example input images results c omparison with Proposed architecture, dataset images are in top row with ground truth, Hover-Net [18] and proposed method predictions in subsequent rows. From Left to right including PanNuke [24], Kumar [25], CPM-15 [27], TNBC [17], CPM-17 [27], CryoNuSeg [26] and CoNSEP [18]

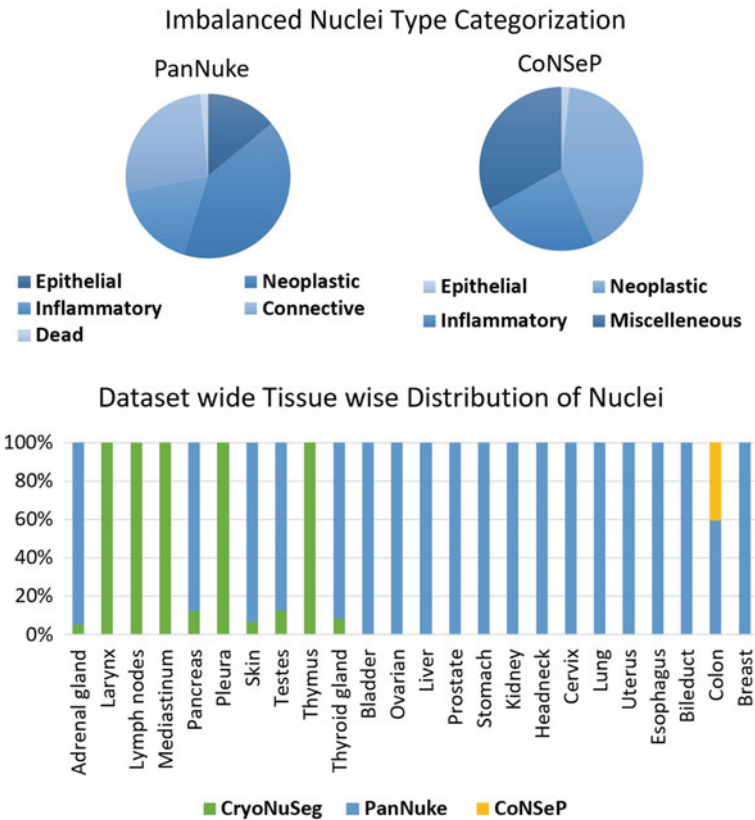


**Fig. 4** Instance segmentation Precision (Pr), Recall (Re), Accuracy, Panoptic Quality (PQ) and Dice Similarity Score (DSC) curves for all seven nuclei datasets including CPM-17, PanNuke, CoNSeP, Kumar, TNBC, CPM-15 and CryoNuSeg datasets

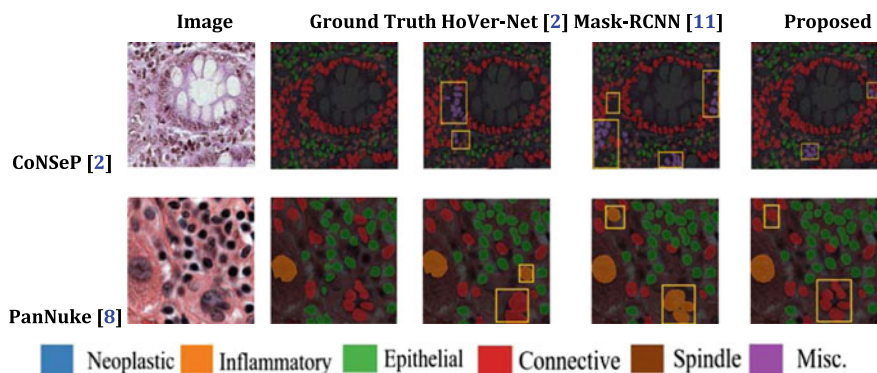
**Table 3** Classification results comparison of proposed architecture with baseline networks for PanNuke and ConSeP dataset in terms of average Panoptic Quality(PQ) and F1 Score

Datasets	Methods	mPQ	$Fe_c$	$Fi_c$	$Fs_c$	$Fm_c$	
CoNSeP	Mask-RCNN [28]	0.450	0.595	0.590	0.420	0.098	
	HoVer-Net [18]	0.516	0.635	<b>0.631</b>	0.566	0.426	
	Proposed	<b>0.55</b>	<b>0.65</b>	0.61	<b>0.58</b>	<b>0.43</b>	
Datasets	Methods	mPQ	$PQ_{ec}$	$PQ_{ic}$	$PQ_{nc}$	$PQ_{cc}$	$PQ_{dc}$
PanNuke	Mask-RCNN [28]	0.37	0.40	0.29	0.47	0.3	0.06
	HoVer-Net [18]	0.46	0.49	0.41	0.55	0.38	0.14
	Proposed	<b>0.48</b>	<b>0.57</b>	<b>0.43</b>	<b>0.57</b>	<b>0.41</b>	<b>0.16</b>

The bold values indicates the best performance result in comparison to all model results



**Fig. 5** Categories distribution in PanNuke and CoNSeP indicates large class imbalance in majority of Publicly available datasets. From sunburst plots we can visualize, high ratio of Neoplastic, Epithelial and Inflammatory classes while rest of the classes constitute extremely less ratio of entire dataset. Tissue wise data distribution of the nuclei categories in publicly available datasets



**Fig. 6** Comparative results visualization for nuclear classification on the CoNSeP and PanNuke datasets

## 5 Conclusion

In this paper, we have introduced an orientation based shape estimation model for dual nuclei instance segmentation and classification. NuRISC yields nuclei structural information and occurrence probability using a light weight encoder decoder model along with a compounded loss function that caters huge class imbalance issue via assigning class weights during loss computation. Thus combining object probability rate with shape estimates producing segmented instances and classification masks thus alleviating the weaknesses of heavier models proposed earlier for dual task performance.

## References

1. Fraz, M.M., Khurram, S.A., Graham, S. et al. FABnet: feature attention-based network for simultaneous segmentation of microvessels and nerves in routine histology images of oral cancer. *Neural Comput & Applic* 32, 9915–9928 (2020). <https://doi.org/https://doi.org/10.1007/s00521-019-04516-y>.
2. Shaban, M., Khurram, S.A., Fraz, M.M. et al. A Novel Digital Score for Abundance of Tumour Infiltrating Lymphocytes Predicts Disease Free Survival in Oral Squamous Cell Carcinoma. *Sci Rep* 9, 13341 (2019). <https://doi.org/10.1038/s41598-019-49710-z>.
3. Shaban, Muhammad & Awan, Ruqayya & Fraz, Muhammad & Azam, Ayesha & Tsang, Yee-Wah & Snead, David & Rajpoot, Nasir. (2020). Context-Aware Convolutional Neural Network for Grading of Colorectal Cancer Histology Images. *IEEE Transactions on Medical Imaging*. PP. 1–1. <https://doi.org/10.1109/TMI.2020.2971006>.
4. G. Murtaza Dogar, Muhammad Shahzad, Muhammad Moazam Fraz, Attention augmented distance regression and classification network for nuclei instance segmentation and type classification in histology images, *Biomedical Signal Processing and Control*.
5. Fraz, M.M., Shaban, M., Graham, S., Khurram, S.A., Rajpoot, N.M. (2018). Uncertainty Driven Pooling Network for Microvessel Segmentation in Routine Histology Images. In: , et al. *Computational Pathology and Ophthalmic Medical Image Analysis*. OMIA COMPAY

2018. Lecture Notes in Computer Science(), vol 11039. Springer, Cham. <https://doi.org/10.1007/978-3-030-00949-6-19>.
6. Nasir, E. S., Perviaz, A., & Fraz, M. M. (2022). Nuclei & Glands Instance Segmentation in Histology Images: A Narrative Review. arXiv. <https://doi.org/10.48550/ARXIV.2208.12460>.
  7. Abdolhoseini, Mahmoud & Kluge, Murielle & Walker, Frederick & Johnson, Sarah. (2019). Segmentation of Heavily Clustered Nuclei from Histopathological Images. Scientific Reports. 9. 4551. <https://doi.org/10.1038/s41598-019-38813-2>.
  8. Sajid Javed, Arif Mahmood, Muhammad Moazam Fraz, Navid Alemi Koohbanani, Ksenija Benes, Yee-Wah Tsang, Katherine Hewitt, David Epstein, David Snead, Nasir Rajpoot, Cellular community detection for tissue phenotyping in colorectal cancer histology images, Medical Image Analysis, Volume 63, 2020.
  9. Bashir, Raja Muhammad Saad & Mahmood, Hanya & Shaban, Muhammad & Raza, Shan e Ahmed & Fraz, Muhammad & Khurram, Syed Ali & Rajpoot, Nasir. (2020). Automated grade classification of oral epithelial dysplasia using morphometric analysis of histology images. 38. <https://doi.org/10.1117/12.2549705>.
  10. G. M. Dogar, M. M. Fraz and S. Javed, "Feature Attention Network for Simultaneous Nuclei Instance Segmentation and Classification in Histology Images," 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2), 2021, pp. 1–6, doi: <https://doi.org/10.1109/ICoDT252288.2021.9441474>.
  11. S. N. Rashid, M. M. Fraz and S. Javed, "Multiscale Dilated UNet for Segmentation of Multi-Organ Nuclei in Digital Histology Images," 2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET), 2020, pp. 68–72, doi: <https://doi.org/10.1109/HONET50430.2020.9322833>.
  12. A. Rasool, M. M. Fraz and S. Javed, "Multiscale Unified Network for Simultaneous Segmentation of Nerves and Micro-vessels in Histology Images," 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2), 2021, pp. 1–6, doi: <https://doi.org/10.1109/ICoDT252288.2021.9441509>.
  13. U. schmidt et al, "Cell Detection with Star-Convex Polygons," in Medical Image Computing and Computer Assisted Intervention - MICCAI, doi: <https://doi.org/10.1007/978-3-030-00934-2-30>.
  14. M. A. Nawshad et al., "Attention Based Residual Network for Effective Detection of COVID-19 and Viral Pneumonia," 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2), 2021, pp. 1–7, doi: <https://doi.org/10.1109/ICoDT252288.2021.9441485>.
  15. Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham.
  16. N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane and A. Sethi, "A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology," in IEEE Transactions on Medical Imaging, vol. 36, no. 7, pp. 1550–1560, July 2017, doi: <https://doi.org/10.1109/TMI.2017.2677499>.
  17. Raju, R., Paul, A.M., Asokachandran, V. et al. The Triple-Negative Breast Cancer Database. Breast Cancer Res 16, 490 (2014). <https://doi.org/10.1186/s13058-014-0490-y>.
  18. Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, Nasir Rajpoot, Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images, Medical Image Analysis, Volume 58, 2019, 101563, ISSN 1361–8415, <https://doi.org/10.1016/j.media.2019.101563>. (<https://www.sciencedirect.com/science/article/pii/S1361841519301045>).
  19. Liu, Zuhao & Wang, Huan & Zhang, Shaoting & Wang, Guotai & Qi, Jin. (2020). NAS-SCAM: Neural Architecture Search-Based Spatial and Channel Joint Attention Module for Nuclei Semantic Segmentation and Classification. <https://doi.org/10.1007/978-3-030-59710-8-26>.

20. S. Graham et al., "Lizard: A Large-Scale Dataset for Colonic Nuclear Instance Segmentation and Classification," 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 684–693.
21. R. Verma et al., "MoNuSAC2020: A Multi-Organ Nuclei Segmentation and Classification Challenge," in *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3413–3423, Dec. 2021, doi: <https://doi.org/10.1109/TMI.2021.3085712>.
22. Nabil Ibtehaz, M. Sohel Rahman, MultiResUNet : Rethinking the U-Net architecture for multi-modal biomedical image segmentation, *Neural Networks*, Volume 121, 2020, Pages 74–87, ISSN 0893–6080, <https://doi.org/10.1016/j.neunet.2019.08.025>.
23. Yeung, M., Sala, E., shonlib, C. B., & Rundo, L. (2022). Unified focal loss: Generalising dice and cross entropybased losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95, 102026.
24. Gamper, J., Alemi Koohbanani, N., Benet, K., Khuram, A., Rajpoot, N. (2019). PanNuke: An Open Pan-Cancer Histology Dataset for Nuclei Instance Segmentation and Classification. In: Reyes-Aldasoro, C., Janowczyk, A., Veta, M., Bankhead, P., Sirinukunwattana, K. (eds) *Digital Pathology. ECDP 2019. Lecture Notes in Computer Science()*, vol 11435. Springer, Cham. <https://doi.org/10.1007/978-3-030-23937-4-2>.
25. N. Kumar et al., "A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology," in *IEEE Transactions on Medical Imaging*.
26. Amirreza Mahbod, Gerald Schaefer, Benjamin Bancher, Christine L'ow, Georg Dorffner, Rupert Ecker, Isabella Ellinger, CryoNuSeg: A dataset for nuclei instance segmentation of cryosectioned H&E-stained histological images, *Computers in Biology and Medicine*.
27. Quoc Dang et al., "Methods for segmentation and classification of digital microscopy tissue images," *Frontiers in bioengineering and biotechnology*, vol. 7, pp. 53, 2019.
28. K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988, doi: <https://doi.org/10.1109/ICCV.2017.322>. doi:<https://doi.org/10.1109/CVPR.2016.90>.

# A Semi-supervised Framework for Automatic Pixel-Wise Breast Cancer Grading of Histological Images



Kenglung Chang, Yanyuet Man, and Hailong Yao

**Abstract** Throughout the world, breast cancer is one of the leading causes of female death. Recently, deep learning methods are developed to automatically grade breast cancer of histological slides. However, the performance of existing deep learning models is limited due to the lack of large annotated biomedical datasets. One promising way to relieve the annotating burden is to leverage the unannotated datasets to enhance the trained model. In this paper, we first apply active learning method in breast cancer grading, and propose a semi-supervised framework based on expectation maximization (EM) model. The proposed EM approach is based on the collaborative filtering among the annotated and unannotated datasets. The collaborative filtering method effectively extracts useful and credible datasets from the unannotated images. Results of pixel-wise prediction of whole-slide images (WSI) demonstrate that the proposed method not only outperforms state-of-art methods, but also significantly reduces the annotation cost by over 70%.

**Keywords** Semi-supervised learning · Deep learning · Breast cancer grading · Expectation maximization model

## 1 Introduction

Breast cancer is the most commonly diagnosed cancer for women, which is estimated to account for 30% of new cancer diagnoses and 15% of cancer deaths in the United States [1]. Early and precise diagnosis of breast cancer is crucial to improve the

---

Kenglung Chang and Yanyuet Man are equal contributions.

This work was supported by the Natural Science Foundation of Beijing, China (Grant No. 7202098).

---

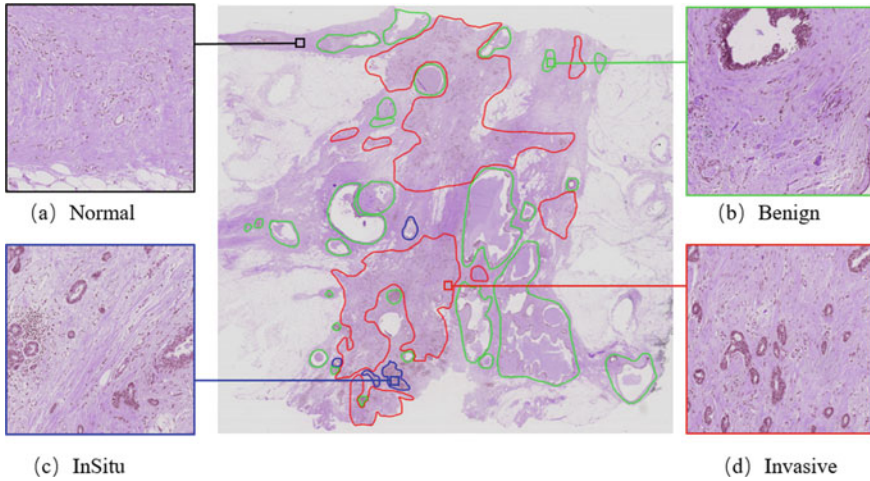
K. Chang · H. Yao (✉)

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China  
e-mail: [hailongyao@tsinghua.edu.cn](mailto:hailongyao@tsinghua.edu.cn)

Y. Man

Tencent AI Lab, Shenzhen, 518057, China





**Fig. 1** An annotated whole slide image: The patches framed in red contour are invasive cancer. Those framed in blue contour are In-Situ cancer. Those framed in green contour are Benign. The rest part of the slide is normal

survival rate of patients [2]. Microscopic examination of stained tissue sections is among the most accurate methods of diagnosing and classifying cancer. The cancer effects can be observed in WSIs in the cellular and tissue levels.

Figure 1 shows an example of the cancerous cell's distribution, which are classified into four categories, i.e., invasive cancer, In-Situ cancer, Benign, and normal. Recently, many computer-aided systems utilize deep learning models to improve the classification consistency and accuracy [3–5]. However, robust deep learning models require large annotated datasets, which are costly to produce especially for medical images. Recent studies integrate active learning with deep learning, which utilize unannotated data to improve the performance of deep learning model [6–9]. Yang et al. applied active learning method on fully convolutional network (FCN) to select the most representative and uncertain areas for annotation [10]. One of the drawbacks is that the FCN cannot be applied to high resolution images, such as WSIs. And it is difficult to acquire iterative annotation on the high-resolution WSIs. Generative Adversarial Network (GAN) is widely applied to generate realistic images, which overcomes the limitations of small training datasets. Mahapatra et al. applies conditional generative adversarial networks (cGANs) to generate informative and realistic chest X-ray images, which enlarge the training datasets [11].

However, GAN generates subtle artifacts on the original images, which could substantially alter the features of cells and tissues, and thus further mislead the model and affect the convergence of parameters. Existing methods fail to provide an efficient solution for automatically grading breast cancer on limited annotated WSIs.

In this paper, we present a new semi-supervised framework based on expectation maximization (EM). We leverage unannotated WSIs to adjust the deep learning model on a limited annotated dataset, reducing the reliance on expensive pixel-wise annotations. The main contributions of the paper are:

- To the best of our knowledge, we first apply active learning method in breast cancer grading, and propose a semi-supervised approach based on EM to effectively reduce annotated dataset for multi-classes pixel-wise breast cancer grading on WSIs.
- We propose a sample selection method based on collaborative filtering, which selects the credible and representative unannotated datasets for enlarging the training dataset.
- Using the proposed semi-supervise framework, significantly enhanced performance on pixel-wise prediction of WSIs is achieved with only 30% of annotated dataset.

## 2 Related Work

Recently, many researchers, as well as vendors of WSI scanning equipments, have started to develop automated WSI image analysis methods to assist pathologists in cancer diagnosis. However, WSI images are too large to be directly integrated into the diagnostic process. The WSI images are typically at the level of tens of millions or even hundreds of millions of pixels, which makes it difficult to store, transmit and visualize. Therefore, traditional algorithms cannot directly process the WSI images. Bejnordi et al. proposed an analytical algorithm at pixel level for automatic detection of ductal carcinoma in situ (DCIS) [12], which detects DCIS across the WSI and differentiates DCIS from good tissue. Balazsi et al. proposed a solution for automatically detecting regions expressing invasive ductal breast carcinomas (IDBC) in images of microscopic tissue or whole digital slides [13]. The proposed method first tessellated whole digital slides. Then image features were extracted and presented to a random forest classifier, which confirms whether each region was cancerous. Cruz et al. proposed a machine learning approach for automatic detection and visual analysis of invasive ductal carcinoma (IDC) tissue regions in whole slide images (WSI) of breast cancer [14]. The adopted convolutional neural network consists of 3 layers. Due to computational limitations, this model is only used for training the images subsampled by 16 times. Rezaeilouyeh et al. proposed a framework for breast cancer detection and prostate Gleason grading using CNN, which was trained on images along with the magnitude and phase of shearlet coefficients [15]. The framework fed shearlet features along with the original images to the CNN consisting of multiple layers of convolution, max pooling, and fully connected layers.

### 3 Method

#### 3.1 Semi-supervised Learning Framework Based on EM Model

An overview of our semi-supervised learning framework is shown in Fig. 2. In the semi-supervised learning framework, only part of whole slide images is annotated, which is defined as set  $D$ . The label of some slides is unknown, which is defined as set  $U$ . Let  $y_i$  denote the label for red patch  $x_i \in D$ . Let hidden variable  $z_j$  denote the label for patch  $x_j \in U$ . We initialize the CNN model on  $D$  and update the model parameter to  $\theta^0$ . We apply initial CNN model to produce the probability map  $P(z_j|x_j)$  of  $x_j \in U$ . The EM algorithm alternates between the E-step for estimating the hidden labels  $z_j$  and the M-step for computing optimal model parameters with maximized  $P(X|Z)$ . The probability map  $P(z_j|x_j)$  is projected to a scaled value between 0 and 1, which is used to generate the consistent heatmap (see Fig. 2c). The fixed vector  $\beta^* = (\beta^1, \beta^2, \beta^3)$  is applied on the heatmap to generate the classmap as shown in Fig. 2d. Next, The most representative and credible patches based on collaborative filtering are selected to train the CNN model in the next iteration.

**Initialization:** Assume the patches are independently and identically distributed (i.i.d.). The initial parameter  $\theta^0$  is obtained from the CNN model, which is trained on annotated dataset  $D$ . Here,  $\theta^0$  is computed as:

$$\begin{aligned} \theta^0 &\leftarrow \arg \max_{\theta} \prod_{x_i \in D} P(x_i, y_i | \theta) \\ &= \arg \max_{\theta} \prod_{x_i \in D} P(y_i | x_i; \theta) P(x_i | \theta) \end{aligned} \quad (1)$$

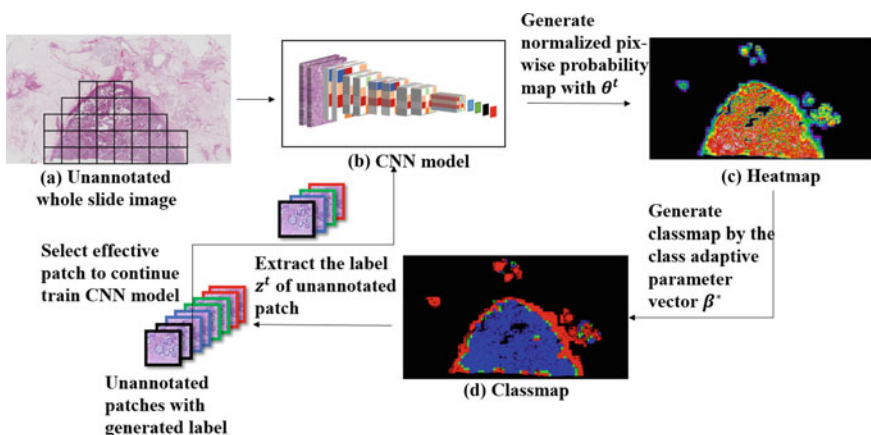


Fig. 2 Overall flow of the EM model

**E-step:** Based on the current parameters  $\theta^t$  at EM iteration  $t$ , we calculate the probability maps  $P(z_j|x_j, \theta^t)$  of unannotated patches, and then re-scale the probability maps to  $P_{norm}(z_j|x_j, \theta^t) \in [0, 1]$ . We generate the class label  $c(x_j)$  based on  $\beta^*$  and then obtain the classmap. The ground truth of the unannotated patches is then extracted as  $c^j$ , and the effective dataset  $E^t$  is selected according to the method described in Sect. 3.2.

**M-step:** The CNN model is retrained on the effective dataset  $E^t$  produced in the E-step. The model parameter  $\theta$  is updated to maximize the likelihood defined in Eq. (2).

$$\begin{aligned} \theta^{t+1} &= \arg \max_{\theta} Q(\theta, \theta^t) \\ K(\theta, \theta^t) &= \prod_{x_i \in D} P(x_i, y_i | \theta) \times \prod_{x_j \in E^t} P(x_j, z_j | \theta) \\ &= \prod_{x_i \in D} P(y_i | x_i; \theta) P(x_i | \theta) \times \prod_{x_j \in E^t} P(z_j | x_j; \theta) P(x_j | \theta) \end{aligned} \quad (2)$$

Assume that  $x_j | \theta$  follows an uniform distribution, we formulate the objective function  $Q(\theta, \theta^t)$  as:

$$\begin{aligned} Q(\theta, \theta^t) &= \log K(\theta, \theta^t) \\ &\propto \sum_{x_i \in D} \log P(y_i | x_i; \theta) + \sum_{x_j \in E^t} \log P(z_j | x_j; \theta) \end{aligned} \quad (3)$$

### 3.2 Patch Selection

Patch selection part can be divided into two stage, Hard Example Mining and Collaborative Filtering.

**Hard Example Mining:** Hard example mining is used in the initialization step to fully exploit the annotated dataset, especially those with wrong classification results. An effective coefficient  $\alpha$  is defined as in Eq. (4). The higher the value of  $\alpha$  is, the harder and more valuable the corresponding patch is for model training.

$$c_k = \arg \max_j p(y_i^{c_j}) \quad \alpha = \|c_k - c_i\| \times P(y_i^{c_k}) \quad (4)$$

Here,  $c_i$  denotes the class label of the patch  $x_i$ , and  $P(y_i^{c_j})$  denotes the probability map.

For the initialization step, we first train our model on 50% of the annotated data. Then, we apply this model on the rest of the data and calculate the effective coefficients. Patches with effective coefficient in the first quintile (top 20%) are selected to retrain the model.

**Algorithm 1** Patch selection method.**Input:**

$\mathcal{U} = \{x_i\}, i \in [1, n]$  (Unannotated dataset)  
 $\mathcal{D} = \{(y_j, L_j)\}, j \in [1, m]$  (Annotated dataset)  
 $\mathcal{M}_t$  (CNN model in iteration  $t$ )  
 $\sigma$  (Similarity threshold for patch selection)

**Output:**

$\mathcal{E}_t$  (Set of unannotated patches in iteration  $t$ )

**Functions:**

*feature*  $\leftarrow F(\mathcal{M}, x)$  {Output  $512 \times 1 \times 1$  feature of  $\mathcal{M}$  given patch  $x$ }  
*prediction*  $\leftarrow P(\mathcal{M}, x)$  {Prediction result of  $\mathcal{M}$  for patch  $x$ }  
 $s \leftarrow \text{sim}(\mathbf{x}, \mathbf{y})$   $\{s = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \times \|\mathbf{y}\|}\}$   
*label*  $\leftarrow \text{argmax}_{\text{index}}(\text{num})$  {Output label with largest number}

**Initialize:**

$\mathcal{E}^t \leftarrow \emptyset$

```

1: for each  $x_i \in \mathcal{U}$  do
2:    $\alpha_i \leftarrow F(\mathcal{M}^t, x_i)$ 
3:    $\text{pred}_i \leftarrow P(\mathcal{M}^t, x_i)$ 
4:   Set num to vector  $[0,0,0,0]$ 
5:   for each  $(y_j, L_j) \in \mathcal{D}$  do
6:      $\gamma_j \leftarrow F(\mathcal{M}^t, y_j)$ 
7:     if  $\text{sim}(\alpha_i, \gamma_j) > \sigma$  then
8:        $\text{num}[L_j] \leftarrow \text{num}[L_j] + 1$ 
9:     end if
10:  end for
11:   $\text{label}_i \leftarrow \text{argmax}(\text{num})$ 
12:  if  $\text{pred}_i = \text{label}_i$  then
13:     $\mathcal{E}^t \leftarrow \mathcal{E}^t \cup x_i$ 
14:  end if
15: end for

```

**Collaborative Filtering:** In the E-step, patches are selected using Algorithm 1. We first apply CNN to extract the features of all patches, and then calculate similarity  $\text{sim}(x_i, y_j)$  between each unannotated patch  $x_i$  and annotated patch  $y_j$ . For each unannotated patch  $x_i$ , we compute the set of annotated patches as  $\{y_j | \text{sim}_{y_j \in D}(x_i, y_j) > t\}$ . Then we apply the majority voting method on the above computed patches to determine the label of the unannotated patch. If the assigned label is consistent with the predicted one by the model, we insert unannotated patch  $x_i$  into  $\mathcal{E}^t$ .

## 4 Experimental Results

### 4.1 Dataset for Training and Validation

The dataset for training and validation is from ICIAR 2018 Challenge<sup>1</sup>.

There are in total 400 images of size  $1536 \times 1536$ , which are obtained from H&E stained breast histology microscopy. The images are labeled into four classes as Normal, Benign, InSitu, and Invasive, respectively. Each class consists of 100 images. Besides, there are 30 whole-slide images, among which 10 images are pixel-wise labeled, and 20 images are not labeled. After foreground extraction and patch cropping, we finally obtained 6389 Normal patches, 695 Benign patches, 369 InSitu patches, and 8182 Invasive patches, where each patch is of size 1536. We randomly selected 30% patches from each class as test dataset, and merged the rest with the above 400 images as the training dataset.

### 4.2 Data Preprocess

**Foreground Patch Extraction on WSI Image:** The high resolution images of WSIs need to be converted into patches for use. However, a large part of WSI is background, which produces uninformative patches in the datasets, and thus should be excluded. The widely used foreground extraction method Otsu fails to extract certain parts of the tissue from the slide for its complexity. It can be clearly recognized from Fig. 3 that the regions circled in red, blue and green bounds are different in their color intensities.

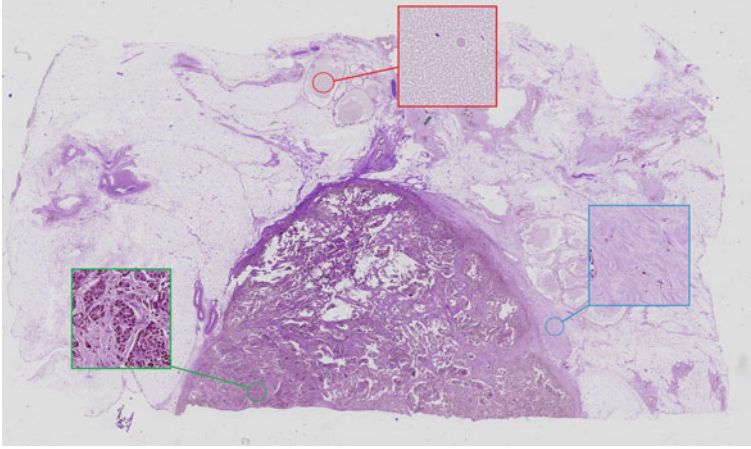
Actually, after converting this image into gray image, the intensity of the pixels circled in red bound is on average 0.76, in blue bound is 0.72, but in green bound is only 0.50, yet the intensity in background is on average 0.86. The high variance in foreground may well reduces the bimodality of the intensity distribution and leads the Otsu method to output the wrong threshold. In the above example, the output threshold by Otsu method (implemented by scikit-mage) is 0.68, which mistakenly classify the region in red circle and blue circle to be background, which in fact are valuable regions containing candidate tissues.

Our adopted method can tackle this problem by concentrating on the relative difference between pixels instead of focusing on the global distribution of intensity. The difference between foreground pixels and background pixels in RGB color space can easily be detected regardless of the high variance in foreground pixels.

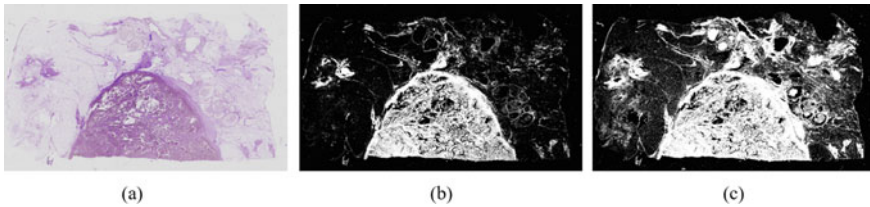
Specifically, we adopt the graph-based image segmentation method in [16] for foreground extraction. For a given slide, we construct an undirected graph  $G(V, E)$ . In  $G$ , each node  $v_{i,j} \in V$  corresponds to a pixel. The edge set  $E = \{(v_{i,j}, v_{i+1,j}), (v_{i,j}, v_{i,j+1})\}$  correspond to the connection between adjacent pixels.

---

<sup>1</sup> <https://iciar2018-challenge.grand-challenge.org/>.



**Fig. 3** An example of whole slide image: The patches in the region circled by red, blue and green boundaries are all foreground patches, whereas Otsu method fails to extract some of them as foreground



**Fig. 4** The foreground extraction of a WSI: **a** The original WSI, **b** Otsu method [18], and **c** our method

We set the edge weight to be  $W(v_{i,j}, v_{i+1,j}) = \|v_{i,j} - v_{i+1,j}\|$ . Then we compute the minimum spanning tree  $T$  using Kruskal's algorithm [17], and delete the edges in  $T$  whose weights are greater than a prespecified threshold (100 in the experiments).

The deletion of these edges produces a forest, i.e., a set of sub-trees (e.g.,  $T_1, T_2, \dots, T_n$ ). Now we compute the average RGB values for the sub-trees (e.g.,  $RGB(T_1), RGB(T_2), \dots, RGB(T_n)$ ). Among the computed average RGB values, assume the maximum value is  $u$ . Then all the sub-trees with average RGB value greater than  $u - 45$  are set as background. Then the foreground mask is obtained as shown in Fig. 4. According to the foreground mask, we crop the WSI into patches with 50% overlap, where each patch consists of  $1536 \times 1536$  pixels. Patches with less than 40% foreground pixels are considered to be background, which are not used for classification.

**Patch Label Extraction:** We assign the label of the patches according to the ground-truth contour of WSI. In most cases, the label of the patch is obtained according to the type of cancer (Benign, In-Situ or Invasive) with the largest area in the patch. However, there are two special cases as follows:

- If the cancer area in a patch is less than one-third of the whole patch area, this patch is labeled as normal.
- If there are two or more types of cancer in a patch, and the corresponding tissue areas are both greater than one-third of the whole patch area, this patch is considered to be a noisy patch and discarded. In the experiments, the number of such patches is very few.

In Kwok’s work in [4], the class value of a patch is the mean of the class values of all pixels in the patch. Our experiments show that the above method tends to generate wrong labels, which disrupt the learning process. For example, when half of a patch contains In-Situ areas and the rest is normal, Kwok’s method labels this patch as benign even if there are no benign tissues at all. In contrast, our EM-based method effectively avoids the drawbacks of Kwok’s method.

### 4.3 Patch Classification

Our neural network is a fine-tuned vgg19 network with batch normalization. Given the large patches of size  $1536 \times 1536$ , we resize them into  $512 \times 512$ , and then feed them into the network. For adapting to the fully connected layers, we add an average pooling layer, which converts the  $512 \times 16 \times 16$  feature map into a  $512 \times 1 \times 1$  vector. The patch-wise experimental results are summarized in Tabel 1. we first apply active learning method (ALM) to continually finetune the classification model with informative and effective datasets instead of retrain the model with all datasets. ALM-10%, ALM-20%, and ALM-30% refer to different models trained on corresponding portions of annotated datasets. FSL is a model trained on all the annotated dataset. Our-10%, Our-20%, and Our-30% are the proposed EM-based model trained on corresponding portions of annotated datasets, in which situation we select the training dataset randomly. For example, Our-30% denotes our proposed method using 30% of the whole annotated dataset.

From the experiment, FSL obtains 0.76, 0.89 and 0.82 for F1 score, accuracy and precision, respectively. Kwok’s method obtains similar results of 0.63, 0.77, and 0.57. However, with 30% of the whole annoatated dataset, ALM obtains 0.83, 0.90, and 0.74, which outperforms FSL. This can be explained by the exclusion of the uninformative data. In contrast, our proposed method obtains 0.86, 0.91, 0.79 for F1 score, accuracy and precision, respectively. Among the different methods, Our-30% achieved the best results using only 30% of annotated data combined with unannotated data. Moreover, our method significantly reduces the runtime for finetuning the model as in ALM.



#### 4.4 Pixel-Wise Classification on WSI

To achieve pixel-wise classification and visualization, we first construct a heatmap for the given WSI. Specifically, we compute the intensity of each foreground patch as  $I = \theta^T \bar{y}$ , where  $\bar{y}$  is the output from the softmax layer, and  $\theta = (0.1, 0.2, 0.7, 1)$  is the weight for each label. Notice that  $I$  indicates the level of severity, when it's closer to 0, the patch is more likely to be normal, but when it's closer to 1, the patch is more likely to have Invasive cancer. We are now able to generate a heatmap for WSI with the intensity of every pixel set as the intensity of the foreground patch it belongs to. Particularly, if one pixel belongs to the intersecting of several patches, we take the mean of the intensities from all the patches as the pixel-level intensity.

Next, we map the heatmap to classmap by the fixed vector  $\beta^* = (0.1, 0.5, 0.75)$ , where a pixel is classified according to its intensity value as follows: (1)  $[0, 0.1]$  for Normal, (2)  $(0.1, 0.5]$  for Benign, (3)  $(0.5, 0.75]$  for In-Situ, and  $(0.75, 1]$  for Invasive. The pixel-wise classification results are summarized in Table 1 in terms of score metrics defined on ICIAR aiming to penalize more on the predictions that are further from the ground truth. The formula for the score is defined as:

$$s = 1 - \frac{\sum_{i=1}^N |pred_i - gt_i|}{\sum_{i=1}^N dist_i \times mask_i} \quad (5)$$

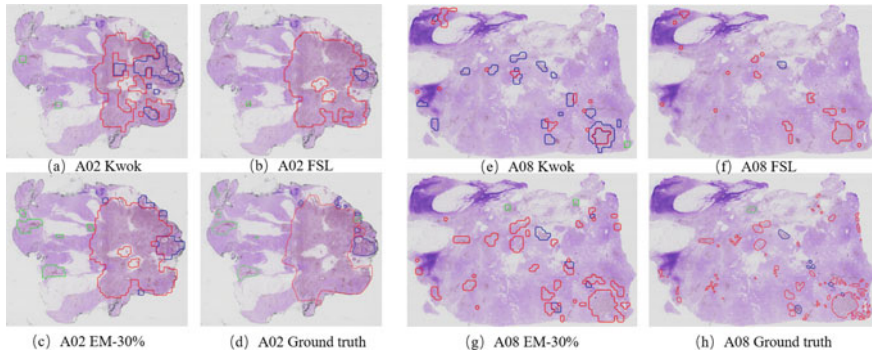
where  $pred$  is the predicted class,  $gt$  is the ground truth class,  $i$  is the index of a pixel in the WSI,  $N$  is the total number of pixels,  $dist_i$  and  $mask_i$  are defined as:

$$dist_i = \max(|gt_i - 0|, |gt_i - 3|) \quad (6)$$

**Table 1** Demographic Prediction performance comparison by three evaluation metrics

Metric	Patch-wise		Pixel-wise	
	Precision	Accuracy	F1-measure	Score metric
Kwok et al.	0.6798	0.8084	0.7391	0.7605
ALM-10 %	0.7205	0.8466	0.8078	0.7186
ALM-20 %	0.7350	0.8684	0.8082	0.7447
ALM-30 %	0.7477	0.9035	0.8303	0.7759
FSL	0.8239	0.8963	0.7698	0.7592
Our-10 %	0.7218	0.8856	0.8054	0.7675
Our-20 %	0.7499	0.8852	0.8048	0.7539
Our-30 %	<b>0.7987</b>	<b>0.9197</b>	<b>0.8623</b>	<b>0.7858</b>
iteration2-Our-30%	0.8293	0.9210	0.8751	0.8027

Kwok [4] is a well-performed Multiclass classification method in whole-slide images which got the first prize in the ICIAR 2018 Challenge. The FSL is a model trained on all the annotated datasets while the Our-xx% are the proposed EM-based model trained on corresponding portions of annotated datasets. Moreover, the ALM-xx% methods are the classic active learning model trained on different portions of annotated datasets. Iterating twice makes sense on the accuracy of multiclass classification. Note that metrics in bold represent the best results with our method in a single iteration



**Fig. 5** The pixel-wise classification results of slides A02 and A08 (Green contour: Benign, Red contour: Invasive, Blue contour: In-Situ, Others: Normal), **a** and **e** give the results of Kwok methods, **b** and **f** give the results of the model trained on full annotated datasets, **c** and **g** give the results of our EM-based method trained on 30% of annotated datasets combined with unannotated datasets, **d** and **h** give the results labeled by pathologist

$$mask_i = 1 - (1 - pred_{i,bin})(1 - gt_{i,bin}) \quad (7)$$

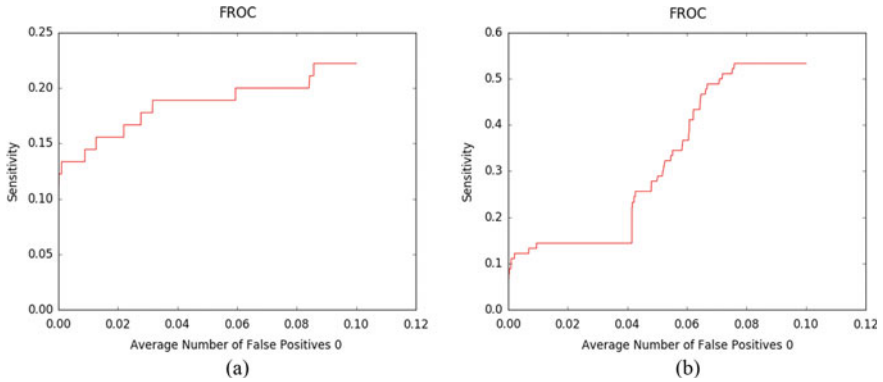
where  $bin$  donates the binarized value, which is 0 if the label is 0 and is 1 if the label is 1, 2 or 3.

Our method with 30% annotated dataset achieves the best performance with a score of 0.785, where the best score of Kwok, FSL, and ALM methods are 0.771, 0.759 and 0.775, respectively.

The three methods Kwok, FSL and ALM perform relatively well in detecting large areas of cancer. However, for small areas of cancer, these methods usually fail. Figure 5 shows an example of invasive tissues in A08 slide, which consists of many small cancer areas. Kwok's method tends to classify small invasive tissues to In-Situ tissues. On the other hand, FSL is unable to recognize lots of small invasive tissues in A02 and A08 slides shown in Fig. 4. In contrast, the proposed EM-based method is able to detect small areas of cancer, which are crucial for correct diagnosis.

#### 4.5 FROC Acceptance

In the medical image processing field, we often use FROC curve instead of ROC curve to validate the effectiveness of a certain model. Figure 6 shows the FROC curves of both the Kwok method and our proposed method. It can be easily seen that Our-30% obtains a much better FROC curve.



**Fig. 6** FROC curve of the Kwok method and our method: **a** Kwok, **b** Our-30%

## 5 Conclusion

In this paper, we have proposed an effective semi-supervised approach based on the EM model, which significantly reduces the reliance on the annotated dataset. Experiment results show that the proposed method achieves remarkable performance with only 30% annotated datasets. Moreover, the proposed method effectively traces the small cancer areas, which is one of the key markers for cancer diagnosis. In the future, more parameters and metrics will be introduced in the system, such as max area of cancer, number of different types of cancer, degree of patient, etc. More prior knowledge will be introduced for generating adaptive parameters in the proposed EM framework.

## References

1. Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019.
2. Walter O’Dell, Cristiane Takita, Katherine Casey-Sawicki, Karen Daily, Coy D Heldermon, and Paul Okunieff. Projected clinical benefit of surveillance imaging for early detection and treatment of breast cancer metastases. *The breast journal*, 25(1):75–79, 2019.
3. Baris Gecer, Selim Aksoy, Ezgi Mercan, Linda G Shapiro, Donald L Weaver, and Joann G Elmore. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. *Pattern recognition*, 84:345–356, 2018.
4. Scotty Kwok. Multiclass classification of breast cancer in whole-slide images. In *International Conference Image Analysis and Recognition*, pages 931–940. Springer, 2018.
5. David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, et al. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136, 2018.

6. Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.
7. Jiming Li. Active learning for hyperspectral image classification with a stacked autoencoders based neural network. In *2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–4. IEEE, 2015.
8. Le Lu, Yefeng Zheng, Gustavo Carneiro, and Lin Yang. Deep learning and convolutional neural networks for medical image computing. *Advances in Computer Vision and Pattern Recognition; Springer: New York, NY, USA*, 2017.
9. Fabian Stark, Caner Hazırbaş, Rudolph Triebel, and Daniel Cremers. Captcha recognition with active deep learning. In *GCPR Workshop on New Challenges in Neural Computation*, volume 10, 2015.
10. Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 399–407, 2017.
11. Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Philippe Thiran, and Mauricio Reyes. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 580–588. Springer, 2018.
12. Van Diest P J et al Bejnordi B E, Veta M. Diagnostic assessment of deeplearning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.
13. Zoroquiain P et al Balazsi M, Blanco P. Invasive ductal breast carcinoma detector that is robust to image magnification in whole digital slides. *Journal of Medical Imaging*, 3(2):027501, 2016.
14. González F et al Cruz-Roa A, Basavanahally A. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *MedicalImaging2014: Digital Pathology*, volume 9041, page 904103. International Society for Optics and Photonics, 2014.
15. Mahoor M H. Rezaeilouyeh H, Mollahosseini A. Microscopic medical image classification framework via deep learning and shearlet transform. *Journal of Medical Imaging*, 3(4):044501, 2016.
16. Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.
17. Joseph B Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956.
18. Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

# Lunatum Prosthetic Replacement: Modeling Based on Volume Rendering of CT Scan Images



Manal Hamda, Btihal El Ghali, Imane Hilal, Omar El Midaoui, Nabil Ngote, Bahia El Abdi, and Kawtar Megdiche

**Abstract** Additive Manufacturing has immersed the medical field, especially in reconstructive surgery, allowing the creation of a 3D model resembling the anatomical structure of interest. Due to Osteonecrosis also referred to as Kienböck's disease; carpal bones especially the lunatum are concerned the most with those technologies especially since a prosthetic replacement is an obligation when it comes to advanced stages of this disease. In this article, we propose a method based on direct 3D reconstruction based on volume rendering directly on patients' medical images (CT scans) to preserve the anatomical shape. For that purpose, we utilized 3D slicer software to create a 3D model based on different cuts of CT scan images. The resulting model was satisfactory, as it was similar to the lunate bone structure preserving all its anatomical characteristics and dimensions. The proposed approach helps in creating a prosthetic replacement with the exact anatomical shape and structure of the bone of interest respecting the dimensions, curves, and facets.

**Keywords** Lunate bone · Lunatum · Kienböck disease · Lunatum replacement · 3D reconstruction · 3D printing technologies · Osteonecrosis

---

M. Hamda (✉) · N. Ngote · B. E. Abdi  
Biomedical Engineering Department, Abulcasis International University of Health Sciences,  
Rabat, Morocco  
e-mail: [Manal.hm1788@gmail.com](mailto:Manal.hm1788@gmail.com)

B. E. Ghali · I. Hilal  
Department of Data and Knowledge Engineering, Information Sciences School, Rabat, Morocco

O. E. Midaoui  
SmartiLAB, Moroccan School of Engineering Sciences (EMSI), Rabat, Morocco

B. E. Ghali · I. Hilal · N. Ngote  
The National School of Mines of Rabat (ENSMR), Rabat, Morocco

B. E. Abdi · K. Megdiche  
Medical Simulation Center, Cheikh Zaid Foundation, Rabat, Morocco

## 1 Introduction

Three-dimensional printing also called additive manufacturing; did take place in the medical field, especially in reconstructive surgery, in other words, a three-dimensional model can be built according to the real anatomical structure using data taken from MRI and CT scans; this reduces the frequency of occurrence and severity of any possible risk or complications arising from prosthetic implantation. This technique is commonly used for total or partial joint replacement or limb salvation surgeries.

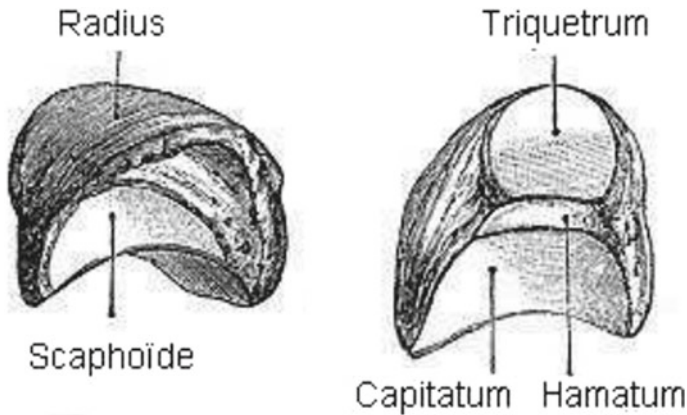
Kienböck's disease (KD) also known as lunatum osteonecrosis has and still causes problems both in its etiology and management. Robert Kienböck, a radiologist was the first to describe "lunatomalacia" clinical characteristics and radiographic aspects in 1910; as a condition caused by the tearing of ligaments and vessels enclosing the semi-lunar bone, which resulted in a fracture and subsequent collapse. In 1843, Peste was the first to provide the characterization of lunatum collapses associated with potential traumatism [1, 2].

In the early stages of Kienböck disease, non-operative measures are used for the treatment of KD, such as anti-inflammatory and painkillers medicaments, physiotherapy, and reducing activities intensity, although, in some cases, surgery is a must to preserve the bone, especially for early-stage osteonecrosis stages [3]. In case of lunatum collapse, prosthetic replacement is recommended to relieve the pain and improve wrist function.

The anatomical structure of the carpal tunnel can be reconstructed and preserved with lunatum arthroplasty, which will eventually reduce the pain and improve wrist mobility [4–6]. However, some patients suffer from mild pain during intense activities [4, 5].

According to the literature; lunatum replacement efficacy had been proved, however, the failure rate cause wasn't clearly explained. The possibility of materials type being the cause is weak; authors have mentioned the material type used for modeling the prosthesis, those were biocompatible and clinically approved [4, 5].

The second possibility is the shape and size; very few literature reviews have mentioned the implant type size wise and manufacturing process [5]; which have high chances chance to be the main cause of failure, especially with the tendency of multiple sizes kits that contain 3–5 lunate prostheses of different sizes and the final shaping retouch done by the surgeon. In this article, we propose a methodology to modulate a prosthetic replacement for the lunatum bone which is a small bone localized in the wrist (carpal bones) based on a direct segmentation after 3D reconstruction of computed tomography images, the goal is to create a prosthetic module similar to the anatomical structure of the bone of interest.



**Fig. 1** Lunatum anatomy and articular surfaces

## 2 Lunatum Anatomy and Associated Problems

The lunatum is a moon-shaped bone located between the scaphoid and triquetrum in the proximal row of carpal bones. The proximal facet that articulates with the radius is convex, whereas the distal facet that articulates with the capitate is concave (Fig. 1) [4].

With no muscular attachments and only a few ligaments to hold it in place [7], the lunatum is more prone to injuries and orthopedic diseases such as fracture, dislocation, and Kienböck disease (KD) which is also known as osteonecrosis or avascular necrosis.

Osteonecrosis remains the principal cause of carpal bones alteration, especially in the lunatum; it is a condition that impacts the blood flow deliberately causing bone collapses; this is mainly caused by vascularization problems arterial disruption to be exact, but may also occur after traumas causing venous congestion with elevated interosseous pressure, it may happen due to a high-intensity traumatic injury or spontaneously [8]. According to previous studies, KD's occurrence is male-dominant and most commonly affects the dominant hand in men aged 20–40 years [9].

## 3 Biomaterial Selection

Biomaterials have an essential role when it comes to implantable prostheses; in fact, they may be the cause of failure rate, their selection is piloted by matching its properties with attended application requirements; in our case, biological requirements have to be specially taken into consideration not to forget the mechanical and physical aspects. Secondary to reactions arising from a foreign body (implants); requirements such as biocompatibility, stress, bioactivity, osteoinduction, and more have become

a necessity for biomaterials when it's come to implantable devices such as prosthetic devices [10].

Starting with Swanson (1970) was the first to introduce lunatum arthroplasty using a silicone rubber implant for the case of KD [11], however, this was abandoned years ago since it causes severe cyst formation due to silicone material with an incidence of 78% [12].

Afterward, in 1984 Titanium lunate arthroplasty (TLA) was introduced to resolve the problems common to silicone lunate implants, TLA clinical outcome was promising with only 20% of failure cases [13]. Titanium and its alloys are initially used for total hip arthroplasty (THA) [14],  $\alpha + \beta$  titanium alloys, such as titanium-6Al-4 V used in THA and TLA [13, 15, 16] has excellent corrosion resistance, low density, and high mechanical strength and biocompatibility with bones [17]. Furthermore, vanadium-free titanium alloys with improved biocompatibility, such as + titanium-6Al-7Nb alloy, have been developed by incorporating biocompatible elements like Niobium [15, 18].

Pyrocarbon is another biomaterial used for lunate arthroplasty; according to a short-term clinical review, Pyrocarbon lunate replacement results were satisfactory for most patients [5, 15]. When compared with titanium prostheses, pyrocarbon is more similar to cortical bone and effectively transfers the load.

Despite their higher tendency to break, pyrocarbon implants are biologically inert and biocompatible, resulting in a lower tendency to tissue reactions when compared to titanium implants [19].

Another biomaterial that has been introduced to orthopedic arthroplasty practices is Polyethylene (PE), which has been widely used in knee arthroplasty since the mid-twentieth century. Polyethylene lunate arthroplasty has a satisfactory outcome [7]. Progress in material manufacturing and processing has led to newer polyethylene with different material properties over the last few decades [20].

Cobalt-chromium (co-Cr) alloys which are supporting metallic materials, were initially used in dentistry, now considered one of the materials most used for THA. Cobalt-chromium alloys characteristics such as strength, corrosion, and wear characteristics make it a great option as an implant material [14].

Zirconia toughened alumina (ZTA or  $Al_2O_3-x\%$  vol  $ZrO_2$ ) which was developed in 2002 is a promising biomaterial used in hip and knee implants [21]; alumina in particles is one of the most successful key materials for THA so far; it has a significant advantage, including good biocompatibility, high mechanical strength, and high fracture resistance. The ceramic material may also have drawbacks in its counterpart, such as inflammatory reactions around the implant [22].

Those are a few examples of the many available clinically proven biomaterials used for implantable prosthesis manufacturing; However, silicon, cobalt-chromium (co-Cr) alloys and Pyrocarbon will be excluded. The first one was abandoned after showing side effects, especially cyst formation, and the cobalt-chromium alloy was labeled as a potential carcinogen according to new European regulation which prohibits the use of implantable medical devices that include more than 0.1% (m/m) cobalt [3, 23, 24], as for Pyrocarbon, only few literature reviews described the clinical outcomes of Pyrocarbon lunate implants [6] which doesn't prove the efficacy,



**Table 1** Biomaterial selection based on standards, biomedical use, printing technics, and weaknesses

Biomaterials	Standards/regulations	Biomedical use	3D printing technic	Weaknesses
Titanium (Ti-6Al-4V) and its alloys	ISO 5832-3:2021	Orthopedic prosthesis mainly in total hip joint	Selective laser melting (SLM)	Osseointegration with the surrounding bone tissue at the initial stage of implantation
		Dental implants		Expensive material
Polyethylene	ISO 5834:2019	Fabrication of porous high-density polyethylene implant for facial and cranial reconstruction	Fused disposition modeling (FDM)	Rare reactions after surgery
		Surgical implants		
Zirconia toughened alumina	ISO 5834:2019	Dental implants	Fused disposition modeling (FDM)	It's may cause reactions such as inflammatory reactions around the implant
		Orthopedic implants		

longevity, and the functionality of this implant [19]. As shown in Table 1 titanium and its alloys require a large investment being it an expensive material as well as its required 3D printing method, in the other hand Zirconia, toughened alumina is less expensive compared to Ti-6Al-4V as a material and the FDM printing method requires less investment compared to SLM however; it is most likely can cause a local reaction (tissues around the implants) which is not ideal. This leads us to polyethylene, this one requires less investment compared to Ti-6Al-4V and rarely has any reaction after surgery, nonetheless Polyethylene is commonly used in invasive medical devices such as intravenous cannulas, tracheal intubation tubes, urinary catheters, and more.

## 4 Material and Methods

### 4.1 Data Acquisition

The utilized database source was computed tomography (CT); a CT scan creates cross-sectional images of the body using rotational x-rays which gives more detailed information than typical X-ray images, it's painless, non-invasive, faster, and less expensive with lower risks than MRI with high accuracy. We based our work on the Data acquired from two adult patients' male and female, for automatic segmentation mask and 3D reconstruction respectively.

## **4.2 Segmentation**

Segmentation is a process commonly used in image processing to divide an image into multiple parts or regions, mostly based on the characteristics of image pixels [25]. The main purpose of this part is to develop an automatic mask for Lunatum segmentation in order to detect the lunatum and automatically segment it from the rest of the hand bones allowing fast localization and diagnosis. For this purpose, we utilized CT scan images of a male patient with a scaphoid fracture to create the model using MATLAB [26].

## **4.3 3D Model Creation Based on Volume Rendering**

Our approach is based on data's direct reconstruction using the volume rendering technique to display CT scan image volumes as 3D objects. After the 3D reconstruction of the patient hand, we segmented the Lunatum bone directly from the 3D model. Therefore, we used The 3D slicer software, an open-source flexible platform designed for image visualization and analysis to create our model [27].

3D slicer is usually utilized for 3D reconstruction of anatomical structures such as bones for studies and teaching proposes [28], or for prosthetic template medialization [29]. In our case, we utilized this platform to create a 3D model of the Lunatum bone, which is relatively a small bone localized in the wrist. The purpose is to test the resulting model structure and compare its similarity with the actual anatomical structure.

## **4.4 3D Printing (Fused Disposition Modeling FDM)**

3D printing is the process to transfer digital data to physical objects, in our case FDM or fused disposition modeling is the method of interest. In this case; the printer machine works disposes of melted filament material (Polyethylene) layer by layer until forming a completed object (Lunatum in our work) [30]. The saved digital design of the Lunatum prosthesis was uploaded into the printer which subsequently automatically transformed it into a physical object, in our case we utilized the stream ultra 3D printer model [31].

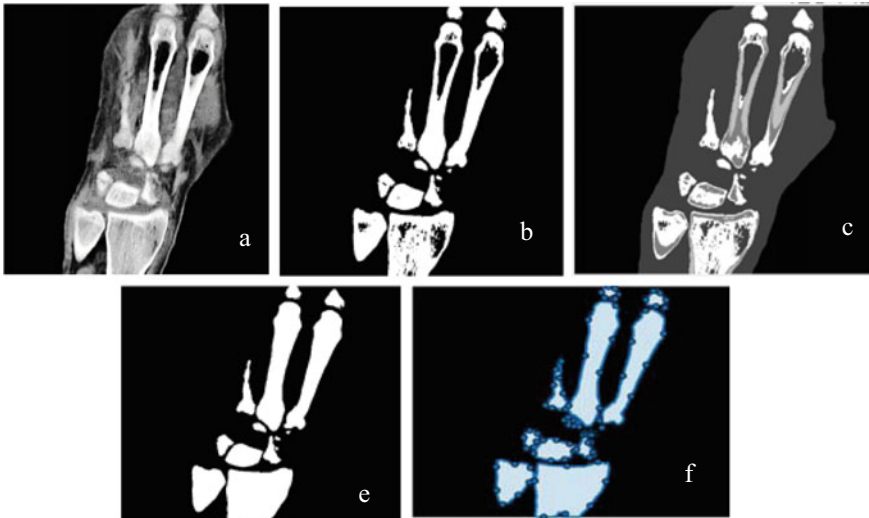
## 5 Results

**Automatic segmentation:** The main goal of this mask is to minimize the consumed time on manual segmentation of the Lunatum bone from the rest of the carpal bones which will allow better diagnosis and classification of KD.

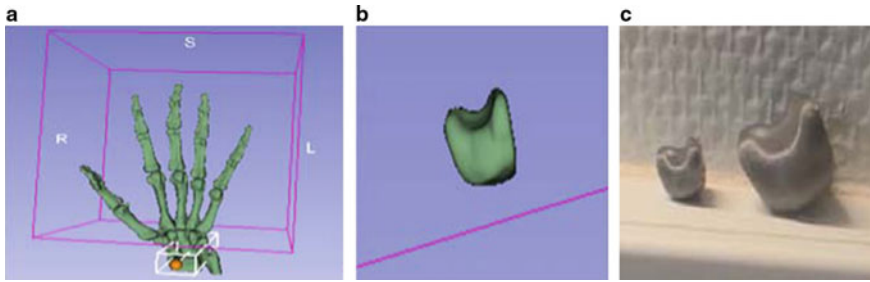
For that, we utilized multiple filters and segmentation technics to reach the desired results; as shown in Fig. 2 image (a) Thresholding results weren't satisfactory, therefore, we used the K-means segmentation method. This one is a partitioned-based algorithm which means dividing analyzed images based on similar features in data to create groups without labels [24]. K-means results were better than Thresholding Fig. 2b, however, the object of interest wasn't fully segmented from the background elements, therefore, we opted for applying K-means two times with different cluster K values ranging from 2 to 5 successively to obtain better results Fig. 2c. The next step was to extract the bone from the binarized image after hole filling Fig. 2d then drew boundaries using the free-hand Roi segmentation method Fig. 2e [32].

**Three-dimensional reconstruction:** our approach reconstructs the 3D model using volume rendering technics Fig. 3a. Therefore, different cuts of CT scan images "sagittal, coronal, and axial" were utilized to visualize the data in 3D volume.

The lunatum 3D model was segmented directly from the previously created 3D hand model using 3D slicer segmentation tools Fig. 3a. The resulting model Fig. 3b was anatomically similar to the patient's bone. The printed model Fig. 3c was similar to the model created using 3D reconstruction Fig. 3b, respecting all its measurements, dimensions, and curves (half-moon shape).



**Fig. 2** Automatic segmentation mask: Thresholding (a) didn't give good results, there for K-means segmentation technic was applied with a cluster ranging from 2 to 5 (b, c), the regions of interest were extracted after holes filling (d) and finally drew



**Fig. 3** Prosthesis modeling process: **a** volume rendering using different cuts of CT scan, **b** segmented 3D Lunatum model, **c** printed lunatums' model based on

## 6 Conclusion

Kienböck disease has always been a clinical challenge in the orthopedics field and bone regeneration. In advanced stages, an arthroplasty is the only solution to decompress the carpal bones, reduce pain, and regain wrist function. Moreover, prosthetic replacement itself is a challenge in terms of compatibility with the bone structure of the wrist, hence the important role of 3D modeling and additive manufacturing. In fact, those technics are mostly involved in prostheses creation and printing.

In this paper, we proposed a simple approach based on 3D reconstruction using volume rendering technics to create a 3D model of the bones' anatomical structure then manually segmented the lunatum bone and printed it using the FDM 3D printing technic. We utilized 3D Slicer an open-access software, simple, and with multiple tools for DICOM data visualization, processing, and 3D reconstruction.

The resulting model was similar to the actual bone in terms of structure, dimensions, curves, size, and facets. Since this method allows direct 3D reconstruction and segmentation from patients' CT scan images, we were able to preserve Lunatums' features, to demonstrate that, we printed the digital model in two different sizes to compare their shape with the original anatomical structure.

The developed automatic segmentation mask allows perfect extraction of the bone from the rest of the carpal bones by drawing ROIs around each bone.

For efficacy evaluation purposes, we applied our segmentation mask to the data of a patient's case with a scaphoid fracture. The main goal was to evaluate the impact of external perturbation (fracture, KD, or dislocation of one of the carpal bones aside from Lunatum) and the possibility to impact the segmentation accuracy.

Since accurate clinical trials require at least 1 year of follow-up to evaluate the outcome of the implant and patient satisfaction in terms of pain relief and movement recovery, we are planning to assess the compatibility of the modulated prosthesis biologically and mechanically in our future works.

For further work, we aim to generalize this technic and create a simple protocol for implanted prosthesis creation, especially for small bones such as carpal bones

and articulation since it allows the preservation of exact anatomical features of the concerned structure.

## References

1. Peltier L. F. (1980). The classic. Concerning traumatic malacia of the lunate and its consequences: degeneration and compression fractures. Privatdozent Dr. Robert Kienböck. Clinical orthopaedics and related research, (149), 4–8.
2. Saunders, B. M., & Lichtman, D. (2011). A classification-based treatment algorithm for Kienböck disease: current and future considerations. *Techniques in hand & upper extremity surgery*, 15(1), 38–40.
3. B. E. Ghali et al., “Modeling and Implementation of a Lunate Implant Based on 3D Reconstruction of CT Scan Images,” 2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS), 2021, pp. 1–7.
4. Ma, Z. J., Liu, Z. F., Shi, Q. S., Li, T., Liu, Z. Y., Yang, Z. Z., Liu, Y. H., Xu, Y. J., Dai, K., Yu, C., Gan, Y. K., & Wang, J. W. (2020). Varisized 3D-Printed Lunate for Kienböck’s Disease in Different Stages: Preliminary Results. *Orthopaedic surgery*, 12(3), 792–801.
5. Xie, M. M., Tang, K. L., & Yuan, C. S. (2018). 3D printing lunate prosthesis for stage IIIc Kienböck’s disease: a case report. *Archives of orthopaedic and trauma surgery*, 138(4), 447–451.
6. Zijlker, H., Fakkert, R., Rijn, J.V., & Beumer, A. (2019). The Short-Term Results of Pyrocarbon Lunate Implants in Patients with Advanced Kienböck’s Disease. *Integrative Journal of Orthopaedics and Traumatology*.
7. Lewis, O.J., Hamshere, R.J., & Bucknill, T. (1970). The anatomy of the wrist joint. *Journal of anatomy*, 106 Pt 3, 539–52.
8. L. Amsallem, J. Serane, D. Zbili, “Idiopathic bilateral lunate and triquetrum avascular necrosis: a case report,” *Hand Surgery and Rehabilitation*, Vol. 35(5), pp. 367–370, 216.
9. Emmanuel J. Camus , LucVan Overstraeten, Kienböck’s disease in 2021, *Orthopaedics & Traumatology: Surgery & Research* Volume 108, Issue 1, Supplement, February 2022, 10316.
10. Allizond, V., Comini, S., Cuffini, A. M., & Banche, G. (2022). Current Knowledge on Biomaterials for Orthopedic Applications Modified to Reduce Bacterial Adhesive Ability. *Antibiotics (Basel, Switzerland)*, 11(4), 529.
11. Ham, S. J., Konings, J. G., & Nielsen, H. K. (1990). Lange-termijnresultaten van de Swansonprothese ter behandeling van lunatomalacie [Long-term results of the Swanson prosthesis for the treatment of lunate osteomalacia]. *Nederlands tijdschrift voor geneeskunde*, 134(37), 1796–1800.
12. Pagnotta, A., Molayem, I. 3D Carpal (Hand) Prosthesis. In: Zoccali, C., Ruggieri, P., Benazzo, F. (eds) *3D Printing in Bone Surgery*. Springer, Cham (2022).
13. Viljakka, T., Tallroth, K., & Vastamäki, M. (2018). Long-Term Clinical Outcome After Titanium Lunate Arthroplasty for Kienböck Disease. *The Journal of hand surgery*, 43(10), 945.e1–945.e10.
14. Chang Yong Hu and Taek-Rim Yoon corresponding author “Recent updates for biomaterials used in total hip arthroplasty” *Biomater Res*. 2018; 22: 33.
15. Rossello M. I. (2020). A case of total scaphoid titanium custom-made 3D-printed prostheses with one-year follow-up. *Case reports in plastic surgery & hand surgery*, 7(1), 7–12.
16. Swanson AB. Flexible implant arthroplasty for arthritic finger joints: rationale, technique, and results of treatment. *The Journal of Bone and Joint surgery. American Volume*. 1972 Apr;54(3):435–455.
17. Yeroushalmi, D., Singh, V., Maher, N., Gabor, J. A., Zuckerman, J. D., & Schwarzkopf, R. (2021). Excellent mid-term outcomes with a hemispheric titanium porous-coated acetabular

- component for total hip arthroplasty: 7–10 year follow-up. *Hip international: the journal of clinical and experimental research on hip pathology and therapy*, 11207000211040181. Advance online publication.
18. Okazaki Y. A new Ti-15Zr-4Nb-4Ta alloy for medical applications. *Current Opinion in Solid State Materials Science*. 2001; 5:45–53.
  19. Cook, S. D., Beckenbaugh, R. D., Redondo, J., Popich, L. S., Klawitter, J. J., & Linscheid, R. L. (1999). Long-term follow-up of pyrolytic carbon metacarpophalangeal implants. *The Journal of bone and joint surgery. American volume*, 81(5), 635–648.
  20. Chakrabarty, G., Vashishtha, M., & Leeder, D. (2015). Polyethylene in knee arthroplasty: A review. *Journal of clinical orthopaedics and trauma*, 6(2), 108–112.
  21. Besim Ben-Nissan, Sophie Cazalbou, Andy H. Choi “Encyclopedia of Biomedical Engineering” 2019, Pages 16–33.
  22. J. Geringer, L. Imbert, K. Kim “Computational modeling of hip implants” *Computational Modelling of Biomechanics and Biotribology in the Musculoskeletal System Biomaterials and Tissues* 2014, Pages 389–416.
  23. Official Journal of the European Union, “L 117 of 5.5.2017. Regulation (EU) 2017/745 of the European Parliament and of the Council of April 5, 2017 relating to medical devices,” April 2017.
  24. Nixon, M.S., & Aguado, A.S. (2002). *Feature Extraction and Image Processing*.
  25. Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15, 29.
  26. Agrawal, P., Shrivastava, S.K., & Limaye, S.S. (2010). MATLAB implementation of image segmentation algorithms. *2010 3rd International Conference on Computer Science and Information Technology*, 3, 427–431.
  27. Pieper, S.D., Halle, M.W., & Kikinis, R. (2004). 3D Slicer. *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*, 632–635 Vol. 1. (IEEE Cat No. 04EX821), 2004, pp. 632–635 Vol 1.
  28. Pujol, S., Baldwin, M., Nassiri, J., Kikinis, R., & Shaffer, K. (2016). Using 3D Modeling Techniques to Enhance Teaching of Difficult Anatomical Concepts. *Academic radiology*, 23(4), 507–516.
  29. Farook, T. H., Jamayet, N. B., Abdullah, J. Y., Asif, J. A., Rajion, Z. A., & Alam, M. K. (2020). Designing 3D prosthetic templates for maxillofacial defect rehabilitation: A comparative analysis of different virtual workflows. *Computers in biology and medicine*, 118, 103646.
  30. Imprimante 3D vomemic.com. <https://imprimante-3d-volumic.com/nouveautae-2018-stream-ultra/>; last accessed 2022/10/09.
  31. Aimar, A., Palermo, A., & Innocenti, B. (2019). The Role of 3D Printing in Medical Applications: A State of the Art. *Journal of healthcare engineering*, 2019, 5340616.
  32. Barry, C. D., Allott, C. P., John, N. W., Mellor, P. M., Arundel, P. A., Thomson, D. S., & Waterton, J. C. (1997). Three-dimensional freehand ultrasound: image reconstruction and volume analysis. *Ultrasound in medicine & biology*, 23(8), 1209–1224.

# Augmented Reality Applications for Image-Guided Robotic Interventions Using Deep Learning Algorithms



Jenna Seetohul, Mahmood Shafiee, and Konstantinos Sirlantzis

**Abstract** Significant breakthrough in the field of surgery has seen the integration of augmented reality (AR) in standard robot operations, allowing anatomical objects to be digitalized and overlaid onto a real-life scenario in-situ. This paper provides an overview of the methodology used to reconstruct and register laparoscopic head and neck image sequences for an AR tool. Deep learning (DL) algorithms are designed to strategically place fiducial markers or labels in a dataset, hence enabling a virtual tool path to be set up for guiding the end effector of a robot. We introduce a dataset of 271 images of patients from four different clinics in Quebec with a proven history of head-and-neck cancer. We then propose a marker-based registration method for mapping a trajectory during surgery, utilizing an unsupervised neural network for computing the medical image transformations. During the training stage, we use an optimized convolutional neural network (CNN) which warps a set of labels from the moving image in contrast to their counterparts in the fixed image. To this end, we compare the loss functions between warped moving labels and fixed labels with respect to the ground truth. Finally, we propose a UNet architecture where we measure the accuracies in label localization throughout the test sequences relative to the initial output results. Our experiments showed that the UNet outperformed the initial CNN architecture, with optimum performance outcomes in losses being closer to 1.0.

**Keywords** Augmented reality · Image registration · Path planning · Supervised learning

---

J. Seetohul (✉) · M. Shafiee · K. Sirlantzis  
School of Engineering, University of Kent, Canterbury, UK  
e-mail: [jls56@kent.ac.uk](mailto:jls56@kent.ac.uk)

M. Shafiee  
e-mail: [m.shafiee@kent.ac.uk](mailto:m.shafiee@kent.ac.uk)

K. Sirlantzis  
e-mail: [k.sirlantzis@kent.ac.uk](mailto:k.sirlantzis@kent.ac.uk)

## 1 Introduction

The use of Augmented Reality (AR) in surgery has plummeted over the past decade, with the ability to provide in situ immersive visualization of a surgical scene in the planning stage as well as during the intervention. Since the groundbreaking release of the Microsoft HoloLens in 2016 [1], the way surgeons perform minimally invasive surgeries has evolved, eliminating inherent challenges that narrow port access and lack of depth estimation causes in the operation theatre. In surgical navigation, most anatomical landmarks are generated in high definition within three dimensional workspaces, from acquired preoperative CT or MRI datasets. The virtual model is registered to the surgical site using fiducial markers, by removing the backend scenes and overlaying a 3D image onto a see-through display [2]. To ensure the safety of the patient and successful final outcomes of surgery, this method of 3D image overlay is ideal for planning in a nonstructured environment. By combining AR with image-guided robotic surgery, the areas of interest in the body can be displayed through a visualization device in real time, improving a surgeon's hand-eye coordination when manipulating the robot end effector. Despite the plethora of studies in existing literature, medical image registration for surgical guidance is still confronted with valuable constraints such as accuracy of label correspondence throughout sequences of images, computational burden on processing units depending on the DL architecture as well as external factors such as signal fluctuations, noise, and acquisition settings.

Our proposed method is an extended framework on the use of two deep convolutional neural networks to compare the output of an optimized registration procedure of the head and neck data with an appropriate transformation which converges to a zero value. In a threefold process, we aim to map the warped moving labels to the fixed labels to earmark the danger zones around the brainstem and spinal cord. We then calculate the dissimilarity between the dynamic and static labels in the CT image sequence using a dice scoring system as well as sum-square-difference (SSD) for intensity-based loss. Finally, we show that by performing a linear transformation such as an affine registration on the network using an alternative DL model such as UNet or probabilistic dense displacements, we can achieve greater accuracy as compared to the existing DeepReg architecture. The output from this experiment can eventually be used for rendering an estimated target trajectory.

## 2 Related Work

In this section, we briefly introduce the use of deep learning for medical image registration, as well as the choice of contrasting models after comprehensive study. We then describe the application of such output databases for AR use in surgery.



## ***2.1 Medical Image Registration Based on Deep Learning***

The innate need for precision in surgical image guidance has seen a dramatic increase in research across the academic community, proposing classic DL algorithms of CT/MRI scans for medical image registration. Ronneberger et al. [3] described the use of conventional neural networks (CNNs) such as the U-Net, where spatial transformations are used to two or more images to a coordinate workspace via an encoder-decoder style network; Qi et al. [4] implemented a modified neural network, PointNet to extract point clouds from medical scans for semantic segmentation which are then used for AR visualisation. Jaderberg et al. [5] proposed a method of applying STNs during both rigid and deformable transformations using transform feature maps on a grid generator. Sokooti et al. [6] described another method of registration called Displacement Vector Fields (DVF) acting as the ground truth and utilized the RegNet architecture for registering CT images of the chest. This enabled a higher accuracy generation when using alternative real-life datasets, in line with the conventional B-spline methods. De Vos et al. [7] proposed an unsupervised end-to-end network using CNNs and STNs to register 2D images of the heart. Inspired by dice loss functions for comparing accuracy in training models, authors such as Hering et al. [8] have touched on existing algorithms for fixed to dynamic segmentation mapping whilst combining CNN-based square difference loss and similarity scores. Balakrishnan et al. [9] extended the work on Voxelmorph for calculating the Dice score between fixed and warped moving segmented masks. Hansen et al. [10] found that the PDD-Net architecture provided a 15% increase in accuracy during monomodal CT registration using a combination of probabilistic dense displacements and differentiable mean-field regularization.

## ***2.2 Augmented Reality Based on DL Image Registration***

The application of AR based technology for surgical guidance has become increasingly relevant in clinics. The use of image superposition for pre-planning of complicated surgeries helps clinicians to transfer the reconstructed medical images from the database to the operating room, for increased tool localisation and reduced operating times. Most clinically approved studies use non-invasive fiducial markers displayed through a visor, to track the position of an end-effector with respect to the patient's body using DL algorithms, libraries and software development kits [11]. Jiang et al. [12] used the principle of medical data registration for detecting simple 2D recognizable objects in a workspace using RGB cameras. Ma et al. [13] used preoperative CBCT images for generating a trajectory during dental implant surgery, where the naked-eye 3D reconstructions were superimposed in-situ to form an AR scene around the patient's mandible using matching markers on the patient's body and the matching CT scans. Wang et al. [14] used SDKs as a computing database for tracking 2D and 3D feature coordinates on medical images and create a calibrated coordinate system

between the real scenario and the digital world. Jiang et al. [15] focused on an AR guided navigation platform for dental implant surgery using mesh to point cloud extraction for preoperative image registration, which achieved lower errors and ( $p < 0.05$ ) for the surgery time.

### 3 Methodology

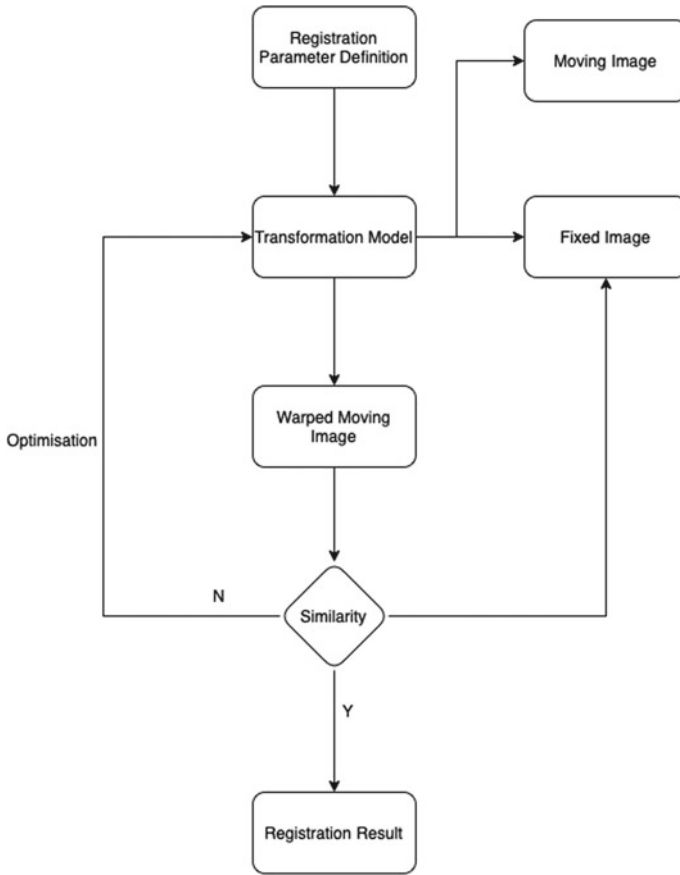
In this study, we propose the use of an existing framework based on deep convolutional neural networks (CNN) for CT scan registration of the head and neck. The unsupervised image registration framework consists of two branches as shown in the block diagram below, one for the moving image,  $M$  and one for the fixed image,  $R$ , each with their associated label. During training, a self-supervised set of labeled data is fed through the neural network, generating a function  $F'$  and is resampled to obtain the warped moving image (Fig. 1).

#### 3.1 Dataset and Implementation Details

CT image reconstructions of the head and neck were generated using a public dataset from The Cancer Imaging Archive. The DeepReg open-source repository is cloned onto the PC terminal to feed input data of 271 test images, each with 37 slices through supervised network. We perform rigid registration of the dataset first where an image coordinate system is initially mapped onto the other to align tissue deformations. This means that only translation and rotation can be performed for target objects to achieve correspondence. Our experiment involves multi-modal registration of real-time CT scans with preoperative ones which will allow for marker-based planning of a trajectory. We aim to use a displacement vector to project the moving coordinates into the static coordinate space. This transformation is characterized as a combination of vectors which allow for all voxels in a CT image to be equalized in a warping procedure. Generally, the voxels within CT images have a wide range of intensity values across their slices which are calculated using intensity histograms. We use measures such as normalized cross-correlation (NCC), mutual information (MI) and basic sum-square-difference (SSD) to measure the common features between moving and fixed images.

#### 3.2 Evaluation Metrics

It is to be noted that the computationally heavy datasets used for image processing require high GPU processing speeds, which generate complex ground truth transformations and therefore DL algorithms such as weakly supervised methods are more



**Fig. 1** Flowchart of registration procedure

suitable for training them. This enables a pair of corresponding moving and fixed labels to be computed, thus extracting the label dissimilarity during the registration. For this experiment, we compare the prediction array to the mask array with the aim of identifying the positive and negative outcomes as well as the mapping results to calculate a loss function for the regions of interest (ROI). The input data includes a probability map from the model, the mask array containing corresponding ground truths and the base threshold predictions. The base image contains 37 slices, with a dimension of  $128 \times 128$  pixels, which is the same for the base label. The outputs include a dense deformation  $\phi$  which has an extra index  $(128, 128, 2)$  because at each pixel, we require a direction vector. The output RGB CT images indicate areas of overlap between masks and predictions. We observe that the labels have transformed from the 0th slice at index 1. Upon magnification, a color-coded outcome chart is used to distinguish among true positives from false positives (FP) and false negatives (FN). The intensity-based loss between the images shown below is calculated

using the mean difference per image tensor. Let dimensions;  $f$  and  $m$  be the fixed and moving image parameters and let  $\varphi$  be the registration workspace that maps the coordinates off onto that of  $m$ . This procedure is depicted as an optimisation problem below:

$$\begin{aligned} \varphi &= \arg_{\varphi} \min L(f, m, \varphi) \\ &= \arg_{\varphi} \min L_{sim}(f, m \cdot \varphi) + \lambda L_{smooth}(\varphi) \end{aligned} \quad (1)$$

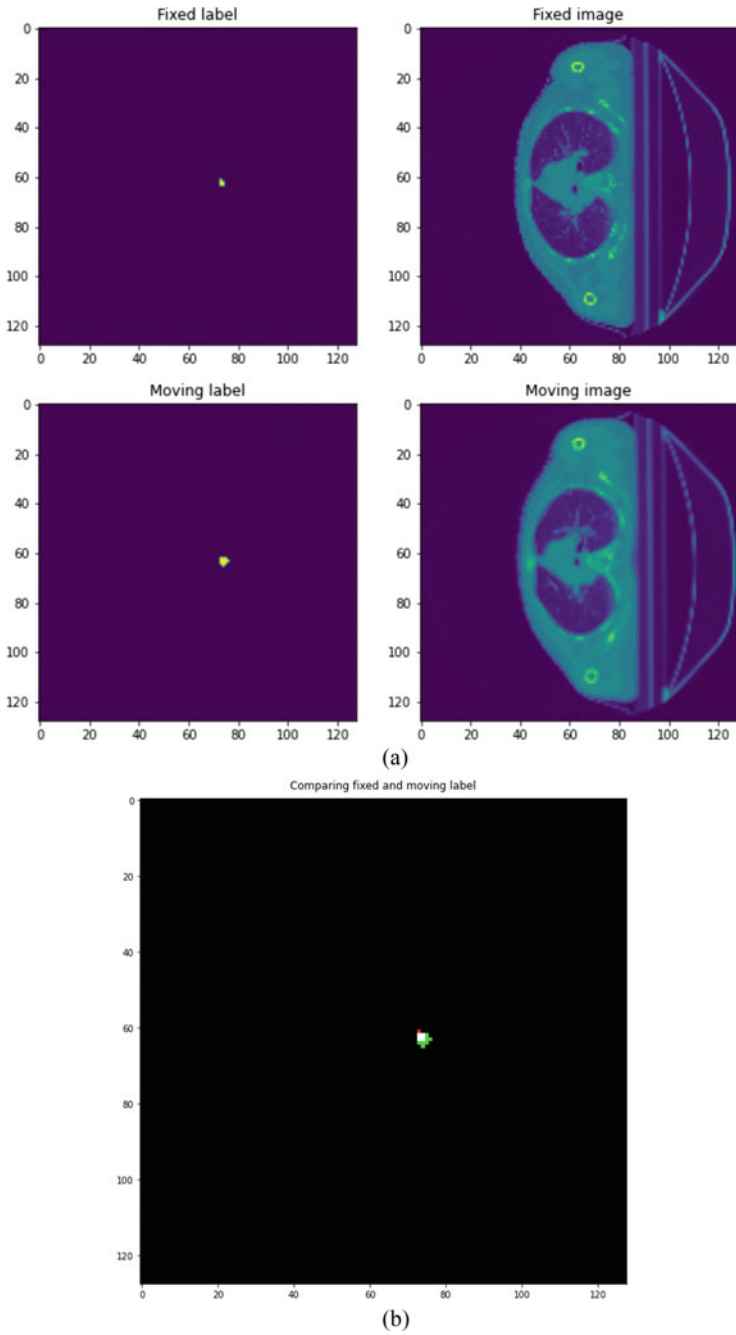
where  $m \cdot \varphi$  represents the warping by  $\varphi$  onto  $m$ , function  $L_{sim}$  represents differences in appearance, and  $L_{smooth}$  shows local spatial variations in  $\varphi$ . The constant  $\lambda$  corresponds to the regularisation trade-off variable. In the process of image registration, we perform voxel-wise correspondence between the fixed and moving datasets whereby we may use affine or non-rigid transformations depending on the degrees of freedom. The function below

$$\mu = \min L(T_{\mu}; I_f, I_m) \quad (2)$$

describes the optimization problem of registering CT images, where  $T$  is the desired spatial transformation which maps  $m$  onto  $f$  and  $S$  is a measure of dissimilarity between the fixed image and the warped moving image. For our experiment, we chose a  $3 \times 4$  affine transformation matrix which is used to visualize the data registration on the fixed images, and then analyze the displacements of consecutive pixels in labels from the test sequence. This means that the straight and parallel lines in the image remain intact but may be translated with a slight change of angle. We found that some of the labels that appeared in the fixed and warped moving images had moved across the slices in the sequence and therefore disappeared from the original moving image. The same process applied for labels in the original moving images disappeared from the fixed and warped moving images, which proved that the warping process was successful (Fig. 2).

### 3.2.1 Control Experiment Using U-Net Module

We focus on the use of supervised learning techniques to predict the outcomes of a particular interventional pathway through the brain. In the control experiment, we use a weakly supervised method to compare the fixed images with their moving counterpart. We then apply another CNN architecture from the VoxelMorph library, adapted from the UNet, to compare the accuracy levels in locating labels in the fixed and warped moving segmentations. U-Net (<http://lmb.informatik.uni-freiburg.de/>) is commonly used for image segmentation tasks and provides accurate registration results. It is developed from the FCN network and has multiple features such as enhanced edge detection, minimized information loss, higher background weight amongst others. In this experiment, we describe the network used with an encoder input of size  $16 \times 32 \times 32 \times 32$  but the framework parameters may vary depending



**Fig. 2** **a** Position of moving label with respect to the fixed label in a  $128 \times 128$  pixel graph, **b** a comparison between the positions of the moving label to the fixed label where TP = white, FP = green and FN = red

on the requirements. We apply three dimensional, 32-layer convolutions in both the encoder and decoder stages using a kernel size of 3 and a stride of 2, where each convolution is followed by a LeakyReLU layer. The process starts with a down sampling step through different degrees of convolutions, followed by a series of up sampling steps and concatenations to decode the network size after learning from the encoding stages. Successive layers of the decoder operate on thinner spatial scaling which enables accurate CT image alignment, whereby the softmax function activates the pixels and generates a probability map.

### 3.3 Experiment Results

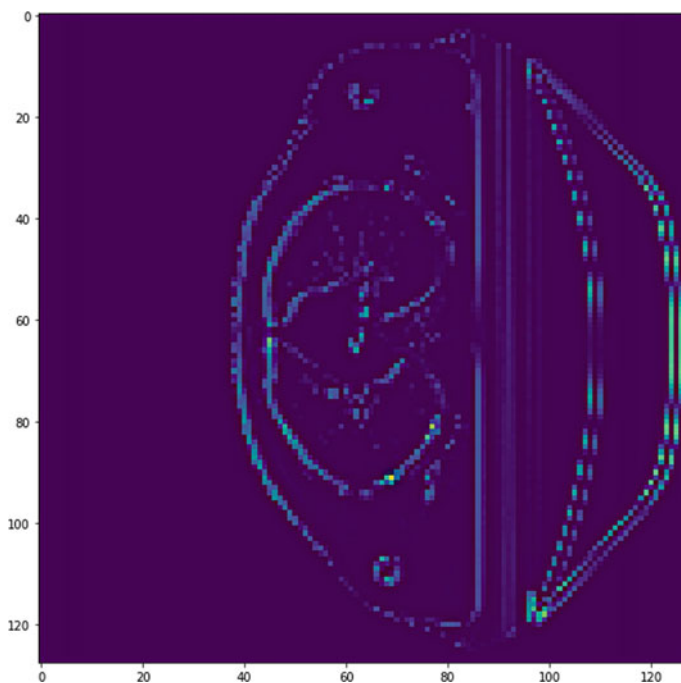
In this section, we present the results of each experiment and attempt to compare the performance based on certain evaluation metrics.

#### 3.3.1 Image Registration

The results of the prediction test (Fig. 3) are shown below in a warped label simulation. We attempt to detect the dissimilarity (SSD) between the fixed label and the moving label by calculating the dice score. In this case, the dice score is 0.517 for 32nd slice of the sequence, where white pixels indicate instances where the model proved that the moving label was in fact located in the same position as the fixed label. The green pixels indicate FP where the moving label was detected in the wrong pixel segment compared to the fixed one and finally, the red FNs indicate a missed segmentation between fixed and moving label. Detecting the image-based loss of each moving tensor or vector compared to the fixed tensor enables us to visualize an average difference between their positions.

#### 3.3.2 Comparison Between U-Net and CNN

In Table 1, we compare the results of the medical image segmentation for the same dataset using both architectures, using metrics such as accuracy, dice scores, SSD and training speeds. The experiment shows clearly that the U-Net and CNN are both suitable for medical image registration. We observed from Table 1 that the U-net network performs better than the original CNN with a dice score of 0.621 on the 32nd slice of the sequence, which was an increase of 10%. Figure 4 shows the difference in intensities and contrasts of the moving label tracked throughout both experiments i.e., the CNN architecture and the U-Net. It is observed that the label appears in most slices in the UNet but appears to fade away during the CNN training, which means that the UNet outperformed the CNN. We compare the accuracy of mapping between fixed and warped moving images, the training speed (s), the F1 score, the SSD value as well as the dice score for each neural network. For control

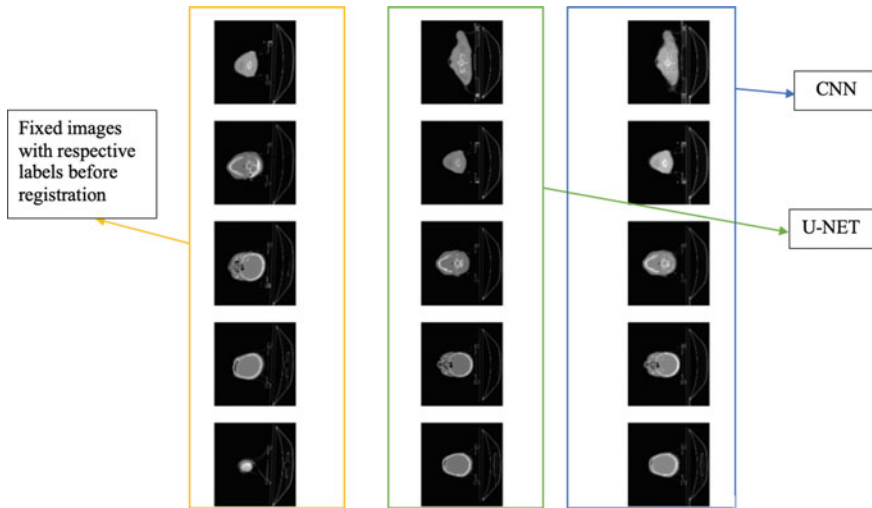


**Fig. 3** CT image reconstruction after performing an SSD between the warped moving image and the fixed image

purposes, we use different neural networks, including the U-Net, CNN, RNN and RegNet to compare the performance of image registration using the aforementioned parameters.

**Table 1** Results of image registration

Parameters	U-Net	CNN	RegNet	RNN
Accuracy	0.71	0.65	0.60	0.77
Training speed (s)	10	30	40	57
F1-score	0.86	0.81	0.751	0.55
SSD	30,450	32,342.5	35,703	19,077
Dice score	0.75	0.51	0.65	0.78



**Fig. 4** Registration of the moving label after being warped throughout the 32-slice sequence during training of the UNet and CNN architecture

## 4 Discussion and Conclusions

Deep learning algorithms have been at the core of medical image registration ever since the concept of surgical visualization emerged in the clinical sector. The precision to which surgeons are now able to perform using image overlays and pre-planning marker-based or marker-less trajectories is a steppingstone towards clinical research in the academic community albeit requiring improvement in the medical image quality for operations which involve morphological and volumetric differences, for example, in the resection of the lung in its deflated state using AR may be impractical since 3D reconstructions are made upon inflated lung CT/MRI scans. Medical image registration requires a high amount of accuracy and efficiency, especially when it comes to complicated cases in surgery where minimal invasion and lower operation times are preferred for quicker convalescence. The use of fiducial markers or “labels” for in-situ AR guidance is an evolving technique which can be used to detect, remove, and alter anatomical landmarks precisely.

This paper uses a variant of the U-net network in parallel with a CNN network from the DeepReg tutorial to analyze and compare the efficacy of label registration on a pre-processed and pre-segmented cancer dataset. The optimized CNN architectures are used for detecting the non-invasive markers throughout the sequence and finally, the segmentation results are compared through relevant evaluation criteria. This method is universal, which means that different datasets can be used for analyzing the performance of both neural networks to obtain an efficient registration technique. However, both methods have their flaws since there may be larger datasets whereby the results are easily influenced by the number of training sets. We are continuously optimizing



the use of neural networks for image and label registration through various supervised learning techniques. The performance of the CNNs can be improved by using an image deformation method, hence reducing dice loss of the labels within a sequence, followed by the generated anatomical path for the surgeon's view.

## References

1. "Holograms replacing cadavers in training for doctors". (Online) Available at: <https://www.theguardian.com/society/2016/nov/17/medical-trainers-look-to-virtual-reality-tech> [Accessed on August 10, 2022].
2. Venkatesan, M.; Mohan, H.; Ryan, J.R.; Schürch, C.M.; Nolan, G.P.; Frakes, D.H.; Coskun, A.F. Virtual and augmented reality for biomedical applications. *Cell Rep. Med.* 2021, 2, 100348. <https://doi.org/10.1016/j.xcrm.2021.100348>.
3. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234–241). Springer, Cham.
4. Qi, K., Yang, H., Li, C., Liu, Z., Wang, M., Liu, Q., & Wang, S. (2019, October). X-net: Brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies. In *International conference on medical image computing and computer-assisted intervention* (pp. 247–255). Springer, Cham.
5. Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. *Advances in neural information processing systems*, 28.
6. Sokooti, H., de Vos, B., Berendsen, F., Ghafoorian, M., Yousefi, S., Lelieveldt, B. P., ... & Staring, M. (2019). 3D convolutional neural networks image registration based on efficient supervised learning from artificial deformations. *arXiv preprint arXiv:1908.10235*.
7. De Vos, B. D., Berendsen, F. F., Viergever, M. A., Sokooti, H., Staring, M., & Išgum, I. (2019). A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52, 128–143.
8. Hering, A., Häger, S., Moltz, J., Lessmann, N., Heldmann, S., & van Ginneken, B. (2021). CNN-based lung CT registration with multiple anatomical constraints. *Medical Image Analysis*, 72, 102139.
9. Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., & Dalca, A. V. (2019). VoxelMorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8), 1788–1800.
10. Hansen, L., & Heinrich, M. P. (2021). GraphRegNet: Deep graph regularisation networks on sparse keypoints for dense registration of 3D lung CTs. *IEEE Transactions on Medical Imaging*, 40(9), 2246–2257.
11. Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., & Yang, X. (2020). Deep learning in medical image registration: a review. *Physics in Medicine & Biology*, 65(20), 20TR01.
12. Jiang, Z., Yin, F. F., Ge, Y., & Ren, L. (2020). A multi-scale framework with unsupervised joint training of convolutional neural networks for pulmonary deformable image registration. *Physics in Medicine & Biology*, 65(1), 015011.
13. Ma, L., Jiang, W., Zhang, B., Qu, X., Ning, G., Zhang, X., & Liao, H. (2019). Augmented reality surgical navigation with accurate CBCT-patient registration for dental implant placement. *Medical & biological engineering & computing*, 57(1), 47–57.
14. Wang, X., Kim, M. J., Love, P. E., & Kang, S. C. (2013). Augmented Reality in built environment: Classification and implications for future research. *Automation in construction*, 32, 1–13.
15. Jiang, W., Ma, L., Zhang, B., Fan, Y., Qu, X., Zhang, X., & Liao, H. (2018). Evaluation of the 3D Augmented Reality-Guided Intraoperative Positioning of Dental Implants in Edentulous Mandibular Models. *International Journal of Oral & Maxillofacial Implants*, 33(6).

# Transfer Learning Based Classification of Diabetic Retinopathy on the Kaggle EyePACS Dataset



Maria Tariq, Vasile Palade, and YingLiang Ma

**Abstract** Severe stages of diabetes can eventually lead to an eye condition called diabetic retinopathy. It is one of the leading causes of temporary visual disability and permanent blindness. There is no cure for this disease other than a proper treatment in the early stages. Five stages of diabetic retinopathy are discussed in this paper that need to be detected followed by a proper treatment. Transfer learning is used to detect the grades of diabetic retinopathy in eye fundus images, without training from scratch. The Kaggle EyePACS dataset is one of the largest datasets available publicly for experimentation. In our work, an extensive study on the Kaggle EyePACS dataset is carried out using the pre-trained models ResNet50 and DenseNet121. The Aptos dataset is also used in comparison with this dataset to examine the performance of the pre-trained models. Different experiments are performed to analyze the images from the different classes in the Kaggle EyePACS dataset. This dataset has significant challenges including image noise, imbalanced classes, and incorrect annotations. Our work highlights potential problems within the dataset and the conflicts between the classes. A clustering technique is used to get informative images from the normal class to improve the model's accuracy to 70%.

**Keywords** Kaggle EyePACS · Pre-trained models · Transfer learning · Deep learning · Fine tuning

---

M. Tariq (✉) · V. Palade · Y. Ma  
Centre for Computational Science and Mathematical Modelling, Coventry University, Coventry,  
UK

e-mail: [tariqm16@uni.coventry.ac.uk](mailto:tariqm16@uni.coventry.ac.uk)

V. Palade

e-mail: [ab5839@coventry.ac.uk](mailto:ab5839@coventry.ac.uk)

Y. Ma

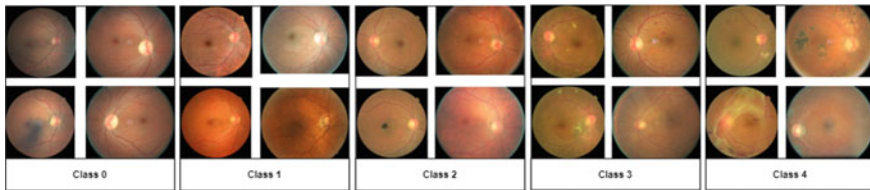
e-mail: [ac7020@coventry.ac.uk](mailto:ac7020@coventry.ac.uk)

# 1 Introduction

Diabetic retinopathy (DR) is an eye complication that can be developed in diabetes, as high blood sugar levels in diabetes damage the eye's retina with time. There are two types of diabetes; Type 1, in which the body does not produce insulin, and Type 2, in which the body produces insulin but does not know how to use it [1]. DR is one of the primary causes of the rise in blindness globally. According to the [1], 422 million adults (aged 20 to 79 years) in 2014 suffered from Type 2 diabetes. Both Type 1 and Type 2 patients are at potential risk of having DR. The population increased to 463 million in 2019 and was predicted to increase to 700 million adults by 2045 [2]. In 2015, there were 2.6 million people that were visually disabled because of DR, and it is expected to rise to 3.2 million by 2020 [3], making DR the leading cause of preventable blindness. The DR is reversible if proper treatment is carried out in the early stages, but there is no permanent cure for this ailment in the later stages [4].

DR can be categorized into five stages; normal, mild, moderate, severe or non-proliferative, and proliferative [5]. It progresses slowly through these stages without proper screening and treatment. During DR, different lesions start appearing gradually in the eye, like microaneurysms in mild DR [6], hemorrhages and exudates in the moderate DR, formation of new blood vessels in non-proliferative DR, and fragile blood vessels and scar tissues in proliferative DR [5]. These lesions slowly distort the retina and further harm the macula. Regular screening and proper treatment after diagnosis are required to prevent this eye-threatening disease [7]. Detection of small lesions is difficult in the initial stages, but it can be very helpful in reducing the risk of severity. The other thing is the correct diagnosis of all five stages of DR to get proper treatment [8]. Human experts and ophthalmologists are available to manually diagnose the signs of DR, which is time-consuming and qualitative. In recent years, much work has been done on the automated detection of DR with the development of relevant technologies [9].

Deep Learning (DL) is an essential tool for processing medical images for classification, object detection [10], and localization [11]. It uses Convolutional Neural Networks (CNNs) to extract features from the images automatically and then distinguishes between images of different classes [12]. In our work, in-depth research on the Kaggle EyePACS dataset is performed to analyze the behavior of the largest available DR dataset. The eye fundus images are first processed through computer vision using different techniques to improve the quality of images. Pre-trained models like ResNet50 and DenseNet121 are trained through transfer learning for multiclass classification to assist human experts in diagnosis. Aptos dataset is used in comparison with the EyePACS dataset to investigate the performance of the developed classification models. In this paper, all experiments are mainly carried out on the Kaggle EyePACS dataset, which has five classes of diabetic retinopathy, as shown in Fig. 1. During classification, many challenges of the EyePACS dataset, such as noise, incorrect labeling, and imbalanced classes, are highlighted. However, this paper focused on the behavior of this dataset, conflicted classes within the dataset, and the potential steps taken to train the model and increase its performance.



**Fig. 1** Images of the five classes of the Kaggle EyePACS dataset

## 2 Related Work

Convolutional neural networks get along well with images but need much time for training [13]. Meanwhile, transfer learning was introduced to achieve better accuracy in less time. It is used to train a previously trained model on an entirely different problem by transferring its learning. The model does not need to be trained from scratch; instead, it learns new data in less time and with reasonable accuracy. GoogLeNet and AlexNet have been used for transfer learning on the Messidor dataset [14]. They have done three experiments with two, three, and four classes to get a test accuracy of 74.5%, 68.8%, and 57.2%, respectively. They have also hypothesized that low accuracy in four classes is due to noise and incorrect labeling [14]. In [8], authors have used Inception-v3 for transfer learning. They have trained their model to do binary classification with a small dataset and managed to get an accuracy of 90.9% with 3.94% of the loss. Inception modules are considered to extract differently sized features of input images in one level of convolution [8]. So, Gulshan et al. have also used Inception-v3 to train their model on binary classification. The model is trained on 0 and 1 as one class and 2, 3, 4 as another class to suggest if the patient needs a referral or not [15].

While working with the Kaggle EyePACS dataset in [5], authors have used data preprocessing and some traditional data augmentation techniques. They have performed two binary classifications; one with healthy (0) and sick (1, 2, 3, 4 classes), and the second with low (0,1) and high (2,3,4 classes). For first classification, they have 94.5% sensitivity and 90.2% specificity. For the second, they have got 98% sensitivity and 94% specificity. For five classes, they have obtained 0.85 of Quadratic Weighted Kappa and 0.74 of F1-score on their test set. In [16], authors have developed a CNN-based system of DR classification using AlexNet, VGG16, and InceptionNet-V3. They have used the Kaggle EyePACS dataset and mentioned the problems within this dataset. The images were handpicked by domain experts to avoid the false labelling of the dataset and achieved a 5-fold cross-validation with the average classification accuracy of 37.43, 50.03 and 63.23% on AlexNet, VGG16, and InceptionNet-V3, respectively. In [17], authors have trained and tested their model on the Kaggle EyePACS dataset. They have achieved a relatively good accuracy of 70%, but on the skewed dataset with the majority of images in class 0. In [18], authors have done a predictive analysis on the Kaggle dataset using transfer learning techniques. It is relatively similar to our work, in which we will perform an intensive analysis of the eye fundus images from the Kaggle EyePACS dataset through different experiments using pre-trained models.

### 3 Pre-trained Models

Two pre-trained models were mainly used for the majority of experiments; ResNet50 and DenseNet121. ResNet50 was introduced with the increased network depth to train more and achieve a reasonable accuracy on the images. We have achieved 92.1% top-5 accuracy and 3.57% top-5 error on ImageNet validation dataset. The architecture of the model is updated and combined with two dense layers for five class classification. DenseNet121 has more depth but slightly less accuracy than Inception-v3, which is 92.3%, and the top-5 error is 7.83% on ImageNet validation dataset. The DenseNet has dense connections between layers, fewer parameters, high accuracy, higher computational efficiency, and memory efficiency. This network advanced the previously developed network ResNet and improves its performance. Like the identity block of ResNet, this network uses a “dense block”. The architecture of the DenseNet121 model is updated, where the base model is combined with the average global pooling layer and dense layer for five class classification in our DR detection problem.

### 4 Dataset

The Kaggle dataset EyePACS was sponsored by the California Healthcare Foundation in 2015, where they launched this competition with the support of a data science team to introduce artificial intelligence in the detection of Diabetic Retinopathy. The images were provided by EyePACS, which is a free platform for retinopathy screening. It consists of 88,696 images, which includes 35,126 images that are annotated for training. Labels are given on the scale of 0–4, which represent the grades of Diabetic Retinopathy. Label 0 shows normal class which includes 25810 images, Label 1 shows mild symptoms of DR which includes 2443 images, Label 2 is moderate DR which includes 5292 images, Label 3 shows symptoms of severe DR and has 873 images, and finally Label 4 shows proliferative DR with 708 images. These grades are given according to the standards of International Clinical Diabetic Retinopathy severity scale by a single specialist. The resolution of images is variable and approximately  $3000 \times 2000$  pixels.

The other dataset we have used is Aptos 2019 (4th Asia Pacific Tele Ophthalmology Society Symposium). APTOS includes 5590 images, 3662 for training and 1928 for testing (Kassani et al., 2019). A clinician has rated each image with the same severity of diabetic retinopathy as in the EyePACS dataset. The number of images is 1805 in the normal class, 370 in the mild class, 999 in the moderate class, 193 in the severe and 295 in the proliferative class. The resolution of images is variable.

## 5 Methodology

In the proposed method, the dataset EyePACS is taken from the Kaggle public repository. This dataset contains images of different resolutions and grades in an excel file. A desktop PC with Nvidia Tesla K80 GPU was used to train the five classes of DR. TensorFlow was used as backend framework.

Data must be preprocessed to remove noise from the dataset and then fed into the pre-trained model for further training. Some preprocessing techniques were applied, which are discussed in this section. The Diabetic Retinopathy images were cropped to the input size of the model, which varies from model to model. For ResNet50, we need  $224 \times 224$  which is quite low, but for DenseNet121, we have changed the input layer of the model to accept the images of custom size  $512 \times 512$ .

### 5.1 Transfer Learning Details

Following are the hyper-parameter details used in these transfer learning experiments.

**Loss Function** Several experiments have been conducted using two different loss functions. Categorical crossentropy loss is used for multiclass classification, but it did not perform well on our dataset due to the imbalanced nature of the dataset or small lesions in the images. The loss function is given below.

$$Loss = - \sum_{i=1}^n y_i \cdot \log(\hat{y}_i)$$

This loss function shows the error between the actual and the predicted output.  $y_i$  is the probability for event  $i$ , which in total equals 1.  $n$  is the number of predictions in the output list.

Sparse Categorical Focal Loss is an extension to categorical crossentropy with the weighting factor  $(1 - \hat{y}_i)$ .  $\gamma$  is the focusing factor used to adjust the rate smoothly. This focal loss works better if the dataset is imbalanced and if there are small lesions within the classes. In this work, focal loss is used with gamma equals to 2. The loss function is below.

$$Loss = \sum_{i=1}^n (1 - \hat{y}_i)^\gamma \cdot y_i \cdot \log(\hat{y}_i)$$

**Early Stopping** Early stopping is used to stop training automatically based on some metric. The metric is usually the validation accuracy or loss that needs to be achieved for the performance evaluation of the model. When this metric stops improving after some epochs, it waits until reaches the value of patience. Patience is the number of epochs without any improvement in the metric. After these epochs, it automatically terminates the training cycle. It increases the model's performance

**Table 1** Conflicting classes

	Classes	Model	Epochs	Accuracy (%)	Class 0	Class 2
Exp 1	0 and 2	ResNet50	120	51	0.61	0.33
	0 and 2	ResNet50	200	50	0.32	0.61
Exp 2	0 and 1	ResNet50	260	50	0.26	0.62
	0 and 1	ResNet50	280	52	0.23	0.65

by avoiding overfitting and saving time. The metric used in this work is validation accuracy and the patience value is 70.

**Optimizer and Learning rate** An optimizer calculates the change after each training cycle and updates the model’s weights. It minimizes the loss value to increase the accuracy. We have tested two optimizers, stochastic gradient descent (SGD) and adam optimizer. SGD is calculated by going through all the training examples. This optimizer did not work for our work; however, the Adam optimizer works well and converges faster for our problem. It has less computation time and needs fewer parameters to tune. The learning rate is set to 0.001, which is considered the best to train the model.

**Model Layers** In the base model, the initial layers of the model have not been trained and frozen to fine-tune the model. Only the last few layers have been trained to extract informative features from the images. After the base model, the global average pooling layer is used to down-sample a patch’s features by taking average values from the feature map. It also reduces the problem of overfitting by learning invariant features. We have used Softmax as an activation function [19], which is used to transform the output before calculating loss in the training cycle. Softmax is used with a dense layer of 5 neurons, and each neuron represents each class.

## 5.2 Training Using Pre-trained Models

EyePACS dataset is the one with the most number of images, but it has a lot of noise, imbalanced classes, and false annotations. We will look into the problems of the Kaggle EyePACS dataset through the conducted experiments.

**Experiment on conflict classes:** This dataset has two major classes, Class 0 and Class 2, with 25810 and 5292 images, respectively. It was considered better to train the majority classes initially and analyze the results. We resized our input images to  $224 \times 224$  for ResNet50 and randomly down-sampled Class 0–5292. The highest accuracy in the two classes was 51%, and the accuracy seemed to be stuck at 50% in the subsequent epochs, which can be seen in Table 1.

**Table 2** Experiments on three classes

	Classes	Model	Resolution of images	Epochs	Accuracy (%)
Exp 3	0, 1 and 2 (Majority), 3 and 4 (Minority)	DenseNet121	224 × 224	80	99.9
Exp 4	0, 3, and 4	DenseNet121	224 × 224	160	66
Exp 5	1, 2, and 3	DenseNet121	224 × 224	80	63.5
Exp 6	1, 3 and 4	DenseNet121	224 × 224	80	69

The same experiment was repeated on Class 0 and Class 1; Class 1 is the next majority class and has 2443 images, so Class 0 was randomly down-sampled to 2443 images. The model responded similarly to Class 0 and Class 1 as the accuracy stuck at 51%. We can say that Class 0 (normal) conflicts with class 1 and class 2. There can be two reasons for this conflict: a mixing between these classes with incorrect annotations, or the model is not good enough to learn small lesions in the initial stages of DR. If we combine conflict classes 0, 1, and 2 as one Majority class and 3 and 4 as Minority class, then it achieves good accuracy, which can be seen in Experiment 3 of Table 2.

**Experiment on Three Classes:** As illustrated in Table 2, it is noticeable that a good accuracy is achieved in Exp 4 and 5. One class is taken from initial grades like 0, 1, and 2, and the other class from severe classes like 3 and 4. It might be due to visible lesions in the images. When the model is trained for minority classes in Exp 6, it can be seen that the DenseNet121 model differentiates well between classes 1, 3, and 4, minority classes. An accuracy of 69% is achieved in 80 epochs.

**Experiment on Five classes:** In Exp 7, DenseNet121 is trained to perform multiclass classification on five classes of DR. The images are resized to a higher resolution of 512 × 512. The accuracy achieved in five classes, with all the traditional image preprocessing techniques, is 48%. The F1-score of each class shows the conflicting nature between classes 0, 1, and 2. In order to defend the ability of the model to learn the lesions, the Aptos dataset was taken to perform multiclass classification on five classes. 80% percent of data was taken from each class for training, and 20% of data was taken for testing. Images were resized to 380 × 380. Our model successfully learned the classes in experiment 8 and achieved a test accuracy of 93% on five classes. The images have good quality, and it is easy to see the small lesions and difference between those classes. Eventually, we can hypothesize that our model is good enough to learn small lesions and differentiate well between five classes. However, this dataset is relatively small, so we cannot standardize this dataset to build a generalized model for DR classification.

In experiment 9, only 700 images were taken from each class to train a Support Vector Machine (SVM). SVM is a non-parametric algorithm implemented to give



**Table 3** Experiments on five classes

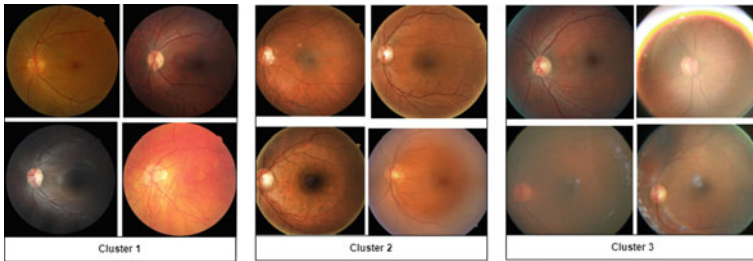
	Model	Dataset	Classes	Accuracy (%)	F1-score
Exp 7	DenseNet121	EyePACS dataset	5	48	Class 0 (0.31) Class 1 (0.48) Class 2 (0.29) Class 3 (0.56) Class 4 (0.68)
Exp 8	DenseNet121	Aptos dataset	5	93	Class 0: 1.00, Class 1: 0.94, Class 2: 0.84, Class 3: 0.95, Class 4: 0.92
Exp 9	Support vector machines	EyePACS dataset	5	52.57	Class 0: 0.35, Class 1: 0.35, Class 2: 0.36, Class 3: 0.79, Class 4: 0.79

the upper estimation of the model's accuracy. The F1-score of the 0, 1, and 2 classes is low, confirming the conflict between these three classes, and our highest accuracy is 52.57%. The five-class classification accuracy is higher on SVM than on neural networks. The results of these experiments can be seen in Table 3.

## 6 Discussion

In this section, the challenges in the Kaggle EyePACS datasets are highlighted and discussed. It has a lot of noise and wrong labeling; however, it is the most used dataset due to its large size. Different image preprocessing techniques have been used to improve noise and increase the quality of images. Data augmentation is implemented during training time to balance the classes of this dataset. Although, the accuracy did not improve as expected. Two pre-trained models, ResNet50 and DenseNet121, were chosen because of their valuable contributions in the medical field to perform multiclass classification. During the training, it was noticed that the model successfully recognized mild classes (0, 1, and 2) from severe classes (3 and 4). However, it did not perform well in differentiating the mild classes (0, 1, and 2) because of the negligible difference between those images. Moreover, class 0 is the shared class that conflicts with both class 1 and class 2, which is why the accuracy got stuck at 50% for these classes. Class 0 is the normal grade class, which holds 70% of the images from the training dataset. So, it can be considered that class 0 has a higher chance of having junk data that requires to be separated.

We have also applied a k-means clustering on class 0 to distribute it into 3 clusters. The purpose of clustering is to separate the informative images from the junk images



**Fig. 2** Images from three different clusters

into one cluster. Each cluster is then investigated with the rest of the classes to see if there is any one cluster that improves the accuracy of the model. After the K-means clustering, the pre-trained model DenseNet121 is used to extract features from all the images of class 0 divided into three clusters. These three clusters are considered class 0 and then trained one by one with other classes 1, 2, 3, and 4. cross-validation is performed to estimate the model’s error. One part of the data is kept for testing from the beginning. The remaining part of the data is used for training the model in a 10-fold cross-validation approach, where weights of the model from each fold training are used to update in the next fold training. Then, the model is tested on the test set partition kept aside from the beginning.

The model’s accuracy increases to 70% on five-class classification when trained on 180 epochs. Our model successfully detects the mild stages of DR, especially class 1 with the small lesion (microaneurysms) of diabetic retinopathy with an F1-score of 0.67. In addition, the detection for the severe stages of DR is also improved with a comparatively better F1-score. The accuracy on the other two clusters is relatively low, which is 42%.

In Fig. 2, we can see some random images from the three clusters 1, 2, and 3. Our model performed well on cluster 2 with 70% accuracy on five classes. It can be seen in Table 4 that the model did well in classifying the four severity classes (1, 2, 3, and 4).

## 7 Conclusion and Future Work

In this paper, we have done a detailed predictive analysis of the Kaggle EyePACS dataset. This dataset is important because it is the largest publicly available dataset with five classes. However, this dataset has many challenges like poor quality, imbalanced classes, and incorrect labeling. In our analysis, we have highlighted the drawbacks of this dataset through different experiments using transfer learning. ResNet50 and DenseNet121 were used as the deep learning models to perform five-class classification. The dataset has three conflict classes, considered to be incorrect-labeled or confused classes; normal, mild, and moderate classes with very few initial symp-

**Table 4** K-means clustering on Class 0

	Cluster	Classes	Accuracy (%)	Model	Epochs	F1-score
Exp 10	Cluster 1	5	42	DenseNet121	180	Class 0: 0.01, Class 1: 0.46, Class 2: 0.22, Class 3: 0.44, Class 4: 0.62
	Cluster 2	5	70	DenseNet121	180	Class 0: 0.13, Class 1: 0.67, Class 2: 0.73, Class 3: 0.85, Class 4: 0.88
	Cluster 3	5	42	DenseNet121	180	Class 0: 0.07, Class 1: 0.48, Class 2: 0.31, Class 3: 0.41, Class 4: 0.57

toms, which is why it is hard to distinguish between them. The Aptos dataset is also used to perform multiclass classification and compared to the EyePACS dataset. However, this dataset is small and insufficient to build a generalized model for DR classification. In future work, it is essential to generate new images for the stages of DR to make a new large dataset that will be good enough to be utilized in real life to help experts in diagnosing Diabetic Retinopathy.

## References

1. Roglic, G.: Who global report on diabetes: A summary. *International Journal of Noncommunicable Diseases* 1(1), 3 (2016)
2. Teo, Z.L., Tham, Y.C., Yu, M., Chee, M.L., Rim, T.H., Cheung, N., Bikbov, M.M., Wang, Y.X., Tang, Y., Lu, Y., et al.: Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology* 128(11), 1580–1591 (2021)
3. Cheloni, R., Gandolfi, S.A., Signorelli, C., Odone, A.: Global prevalence of diabetic retinopathy: protocol for a systematic review and meta-analysis. *BMJ open* 9(3), e022188 (2019)
4. Shibib, L., Al-Qaisi, M., Ahmed, A., Miras, A.D., Nott, D., Pelling, M., Greenwald, S.E., Guess, N.: Reversal and remission of t2dm—an update for practitioners. *Vascular Health and Risk Management* 18, 417 (2022)
5. Islam, S.M.S., Hasan, M.M., Abdullah, S.: Deep learning based early detection and grading of diabetic retinopathy using retinal fundus images. *arXiv preprint arXiv:1812.10595* (2018)
6. Gori, N., Kadakia, H., Kashid, V., Hatode, P.: Detection and analysis of microaneurysm in diabetic retinopathy using fundus image processing. *vol 3*, 907–911 (2017)
7. Cavan, D., Makaroff, L., da Rocha Fernandes, J., Sylvanowicz, M., Ackland, P., Conlon, J., Chaney, D., Malhi, A., Barratt, J.: The diabetic retinopathy barometer study: global perspectives on access to and experiences of diabetic retinopathy screening and treatment. *Diabetes research and clinical practice* 129, 16–24 (2017)

8. Hagos, M.T., Kant, S.: Transfer learning based detection of diabetic retinopathy from small dataset. arXiv preprint [arXiv:1905.07203](https://arxiv.org/abs/1905.07203) (2019)
9. Gao, Z., Li, J., Guo, J., Chen, Y., Yi, Z., Zhong, J.: Diagnosis of diabetic retinopathy using deep neural networks. *IEEE Access* 7, 3360–3370 (2018)
10. Mathe, S., Pirinen, A., Sminchisescu, C.: Reinforcement learning for visual object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2894–2902 (2016)
11. Al, W.A., Yun, I.D.: Partial policy-based reinforcement learning for anatomical landmark localization in 3d medical images. *IEEE transactions on medical imaging* 39(4), 1245–1255 (2019)
12. Sunghheetha, A., Sharma, R.: Design an early detection and classification for diabetic retinopathy by deep feature extraction based convolution neural network. *Journal of Trends in Computer Science and Smart technology (TCSST)* 3(02), 81–94 (2021)
13. Abramoff, M.D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., Folk, J.C., Niemeijer, M.: Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative ophthalmology & visual science* 57(13), 5200–5206 (2016)
14. Lam, C., Yi, D., Guo, M., Lindsey, T.: Automated detection of diabetic retinopathy using deep learning. *AMIA summits on translational science proceedings 2018*, 147 (2018)
15. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316(22), 2402–2410 (2016)
16. Wang, X., Lu, Y., Wang, Y., Chen, W.B.: Diabetic retinopathy stage classification using convolutional neural networks. In: *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. pp. 465–471. IEEE (2018)
17. AATILA, M., LACHGAR, M., HRIMECH, H., KARTIT, A.: Diabetic retinopathy classification using resnet50 and vgg-16 pretrained networks. *International Journal of Computer Engineering and Data Science (IJCEDS)* 1(1), 1–7 (2021)
18. Salvi, R.S., Labhsetwar, S.R., Kolte, P.A., Venkatesh, V.S., Baretto, A.M.: Predictive analysis of diabetic retinopathy with transfer learning. In: *2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE)*. pp. 1–6. IEEE (2021)
19. Sharma, S., Sharma, S., Athaiya, A.: Activation functions in neural networks. *towards data science* 6(12), 310–316 (2017)

# Ex-vivo Evaluation of Newly Formed Bone After Lumbar Interbody Fusion Surgery Using X-ray Micro Computed Tomography



Jakub Laznovsky , Adam Brinek , Tomas Zikmund ,  
and Jozef Kaiser 

**Abstract** Many novel biomaterials are recently investigated for use in spinal fusion surgery, especially in lumbar interbody fusion. The X-ray microCT as a tool is widely used for evaluating how successfully those biomaterials can perform a vertebral fusion. However, the current methodologies of microCT image assessment are based on visual evaluation by the operator. In this paper, we propose a methodology for how such biomaterials can be investigated in pre-clinical studies by investigating fused vertebrae morphology. We utilized microCT scans of pigs' fused vertebrae to develop a fully automatic approach, which can characterize the morphometry of the bone in the fused region. A surface mesh model was created to extract the newly formed bone tissue between fused vertebrae in the microCT data. Extracted bone tissue was consequently evaluated according to the selected morphometric parameters. Characterization of the newly formed bone properties in the intervertebral area can be utilized to evaluate the osteogenesis function of implants used in lumbar interbody fusion surgery.

**Keywords** X-ray micro computed tomography · Intervertebral fusion · Image processing · Lumbar interbody fusion · Quantitative tomography

---

J. Laznovsky (✉) · A. Brinek · T. Zikmund · J. Kaiser  
CEITEC-Central European Institute of Technology, Brno University of Technology, 612 00 Brno,  
Czech Republic  
e-mail: [Jakub.Laznovsky@ceitec.vutbr.cz](mailto:Jakub.Laznovsky@ceitec.vutbr.cz)

A. Brinek  
e-mail: [Adam.Brinek@ceitec.vutbr.cz](mailto:Adam.Brinek@ceitec.vutbr.cz)

T. Zikmund  
e-mail: [Tomas.Zikmund@ceitec.vutbr.cz](mailto:Tomas.Zikmund@ceitec.vutbr.cz)

J. Kaiser  
e-mail: [Jozef.Kaiser@ceitec.vutbr.cz](mailto:Jozef.Kaiser@ceitec.vutbr.cz)

## 1 Introduction

Spinal fusion is a neurosurgical technique that connects two or more vertebrae to prevent a motion between them. This technique is performed for the treatment of various degenerative diseases to relieve back pain and pressure. Since the early 1900s, bone grafts have been used as a source of growth factors to reach a permanent vertebral fusion. The bone graft (autograft) is surgically removed from another part of the patient body, usually from the iliac crest. This method remains a standard up to recent times. Currently, the huge expanse of biomaterials used in medicine brings many new approaches to spinal fusion every year [1–4]. Usage of biomaterials is beneficial in this case due to the possibility of fusion rate regulation and complicated obtaining of the autografts. Evaluation of the vertebral fusion quality in order to evaluate individual biomaterials is therefore fundamental.

Micro Computed Tomography (microCT) plays an important role in the fusion quality assessment. Thanks to the 3D non-destructive visualization and quantitative analysis of Lumbar Interbody Fusion (LIF) location, it is possible to evaluate bone tissue properties. According to previous studies, the accuracy of microCT for bone morphometry is closely correlated with histomorphometric techniques [5–7]. In the case studies, which can proceed *ex-vivo*, the advantage of microCT can be taken. The main benefit of microCT compared to a clinical CT scanner is the spatial resolution of the scan in order of micrometers (dependent on the size of the sample).

In the case of LIF quality assessment using microCT, it is crucial to select an objective and standardized approach for the LIF area analysis. Several automated approaches for the LIF area were already introduced but usually require some enhancement or are suitable for a method other than microCT, especially for clinical applications (plain radiography, clinical CT, magnetic resonance imaging) [8–11]. Another category is visual methods, which are established but depend on the subjective evaluation by the operator [12, 13]. The development of a standardized approach can facilitate the comparison of the vertebral samples, where vertebrae are fused with different types of intervertebral implants, including bone grafts.

In this work, we extended analyses from [14] and analyzed the vertebral samples after LIF in detail using quantitative parameters evaluating the newly formed bone properties. The main motivation is to provide a tool which can easily and objectively analyze LIF area structure, using different biomaterials used for vertebral fusion. Such a methodology can consequently facilitate and accelerate the investigation of biomaterials suitable for vertebral fusion. Automation of this process is crucial, since manual methods are affected by bias caused by the operator.

## 2 MicroCT Bone Tissue Evaluation

The formation of new bone after LIF surgery is possible thanks to the osteogenic, osteoinductive, and osteoconductive properties of fusion materials used for vertebral fusion [15]. Since bone fusion is artificially created, the properties of the newly formed bone tissue may vary from individual to individual. Bone morphometry is able to quantitatively describe the correlation between the growth and development of the examined bone and the type of material used for spinal fusion.

Besides, microCT is capable to evaluate bone samples to study metabolic bone diseases such as osteoporosis and characterize the efficiency of therapies for these degenerative diseases [16]. The main benefit is the non-destructive evaluation of bone fragility, microdamages, and density. Consequently, it is possible to create 3D models of examined bones (vertebrae) for simulations of mechanical stress, and bone fragility induced by loading [17].

There are several morphological parameters that characterize the bone and can be derived directly from the microCT 3D image stack. These parameters are obtained by image-processing methods using various software provided by microCT manufacturers or by applying mathematical methods in a programming environment. There are four basic parameters characterizing the trabecular bone: Trabecular thickness (Tb.Th), separation (Tb.Sp), number (Tb.N), and bone volume fraction (BV/TV) [18]. Mean Tb.Th and Tb.Sp are evaluated using the sphere fitting method, where in the case of Tb.Th the biggest spheres inscribed in the individual parts of the segmented object are considered. In the case of Tb.Sp is the approach similar, but spheres are fitted into the gaps between trabeculae (image background). Individually fitted sphere diameters are consequently averaged to obtain a single representative Tb.Th or Tb.Sp value. BV/TV is based on the ratio of voxels belonging to the bone and to the volume of interest (VOI), and Tb.N can be derived as the proportion of BV/TV and Tb.Th.

Further parameters evaluating trabecular bone are Connectivity Density (Conn.D) and Degree of Anisotropy (DA). Connectivity is designed to estimate the number of connected trabeculae in a trabecular network. The calculation of connectivity is based on the Euler characteristics, which count the number of objects in VOI, the number of marrow cavities surrounded by bone, and the number of connections that must be broken to split the structure into two parts. A more convenient approach is to relate the connectivity to the total volume of VOI and express this parameter as connectivity density [19]. The Degree of Anisotropy describes the orientation of the structural elements in the bone. DA specifies whether the trabeculae have a particular orientation or are arranged randomly. The calculation is based on the mean intercept length from various directions [20].

## 3 Materials and Methods

### 3.1 Datasets Used

In this work, X-ray microCT data of 4-month-old pigs after LIF surgery were used. One-level LIF surgery was conducted on Lumbar 2 and Lumbar 3 (L2–L3) vertebrae. The samples were divided into three groups according to the material for LIF used. A bone autograft from the iliac crest was used in the first group (group A). In the second group (group B) was used a biodegradable nanocomposite implant of biphasic calcium phosphate [2] modified with collagen/oxycellulose biopolymeric foam, enriched with fibroblast growth factor 2 [21]. In the third group (group C), similarly composed biomaterial as in group B was used, but the fibroblast growth factor 2 was substituted by bioactive polyphosphate. All samples were after the LIF surgery fixed with the pedicle screws [22].

The fused vertebrae were surgically removed, wrapped into the plastic foil to avoid samples drying, and scanned on microCT system GE phoenix vltomelx L 240 (Waygate Technologies, USA). The voltage of the scan was 100 kV, the current was 300  $\mu$ A and the X-ray beam was filtered by a 1.5 mm aluminum filter. In total, 2200 projections were captured with the detector exposure time of 400 ms. For more about the samples and their measurement, see [14].

### 3.2 Determination of Volume of Interest

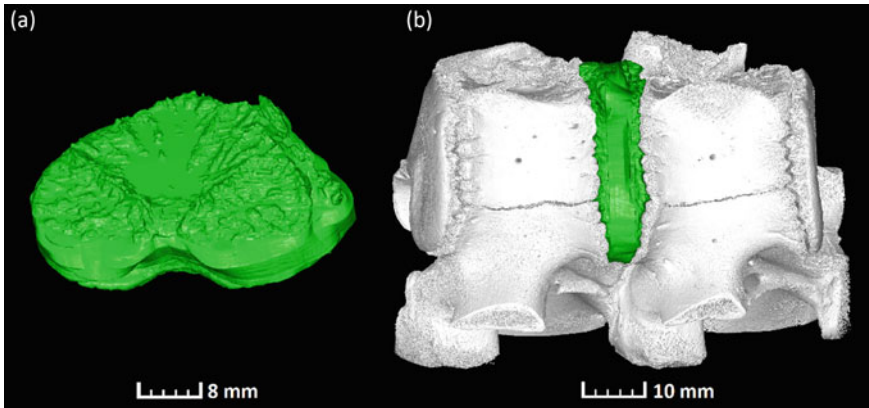
All datasets were firstly registered in the coordinate system according to the top-cranial and bottom-caudal orientation, where the L2 vertebra is located in the upper part of the volume. Consequently, a prepared surface mesh representing the LIF area was fitted on the sample using VG Studio MAX 3.4 (Volume Graphics GmbH, Germany). The manually pre-fitted mesh was consequently automatically registered using the best fit tool. The mesh fitted in the 3D volumetric data created the VOI. VOI was consequently extracted and further analyzed (see Fig. 1b).

Preparation of the mesh representing the LIF area was conducted by manual segmentation of the LIF area in 6 samples. Binary masks were consequently averaged and smoothed using a gaussian filter (see Fig. 1a). This procedure was conducted in Matlab (MathWorks, Inc).

### 3.3 Image Analysis

Evaluation of the newly formed bone in the LIF area is based on the quantification of seven parameters: Trabecular In Growth Ratio (TIGR) acquired from [14], mean





**Fig. 1** **a** Surface mesh representing the LIF area. **b** Mesh registered into the sample in order to extract the LIF area for further evaluation

Tb.Th, mean Tb.Sp, BV/TV, Tb.N, Conn.D and DA. TIGR value represents the ratio between the fused area, and the area of facies intervertebralis.

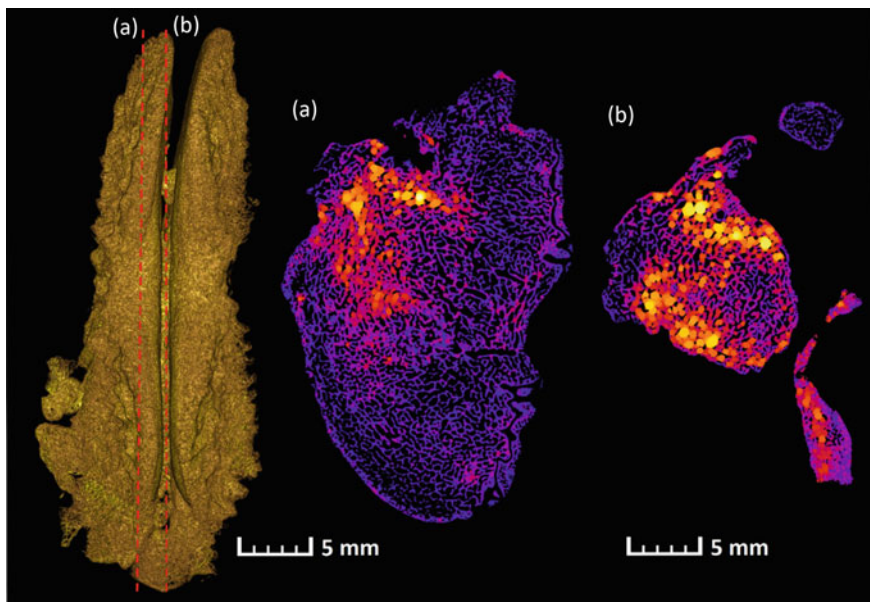
Within a determined VOI was quantified a bone tissue volume (BV), using VG Studio MAX 3.4. The suggested threshold according to the image histogram for advanced surface determination was used. Since the mesh fitted into the LIF area of individual samples always has the same volume, the total volume (TV) used in the BV/TV parameter is determined in advance.

Extracted VOIs of each measured sample were processed in ImageJ software using the BoneJ plugin [23]. Three parameters were quantified by BoneJ: Tb.Th, connectivity, density, and anisotropy. Firstly, the samples were segmented to extract the bone volume. Segmentation proceeded using Otsu thresholding, according to [24]. Consequently, the binary mask was purified to remove all particles. Purification is based on the analysis of connected components and removes all particles surrounding the largest component. Such particles may have been formed by potential noise in the data. Lastly, Tb.Th, Conn.D, and DA were calculated using the BoneJ—see Tb.Th calculation in Fig. 2). Since the anisotropy calculation is a stochastic process, the calculation proceeded three times, and the mean value was chosen as a representative.

Tb.N and Tb.Sp were calculated according to the following equations: (Eqs. 1, 2 respectively):

$$\text{Tb.N} = (\text{BV}/\text{TV})/\text{Tb.Th} \tag{1}$$

$$\text{Tb.Sp} = (1/\text{Tb.N}) - \text{Tb.Th} \tag{2}$$



**Fig. 2** Analysis of trabecular thickness. **a** Cross-section in the location of facies intervertebralis, **b** Cross-section in the area of LIF. Brighter color depicts larger trabeculae

## 4 Results

Evaluation of the newly formed bone is characterized by obtained parameters in Table 1. Samples are divided into three groups. Group A—bone graft, group B and C—different biomaterials (see Sect. 3.1). According to the TIGR<sup>1</sup> value, vertebrae fusion proceeded the best by the samples in group B, taking into account the average value. The standard deviation, on the contrary, is the highest because sample 2 in this group did not fuse at all. The TIGR value coincides with the mean Tb.Th value, which is also the highest in group B, and also has the highest standard deviation.

In the samples where the bone graft was used (group A), trabeculae formed with the greatest distance apart of all groups (mean Tb.Sp = 0.45), but the trabeculae had the highest value of connectivity density ( $4.14 \text{ mm}^{-3}$ ). The evaluation parameters in group C manifest the lowest amount of newly developed bone. This is given by insufficient osteogenesis function of the bioimplant used within this group. Especially the TIGR value and Conn.D parameters indicate the fusion fragility.

The parameters characterizing the morphometry of the newly formed bone do not manifest big differences among individual groups—see graph in Fig. 3. The largest percentage difference is in Tb.N, where group B has the highest amount of trabecular bone. This is related to the small Tb.Sp value in this group and thus the increased BV/TV value. On the contrary, the smallest difference is in the mean Tb.Th value.

<sup>1</sup> Trabecular in Growth Ratio (TIGR) acquired from [14].

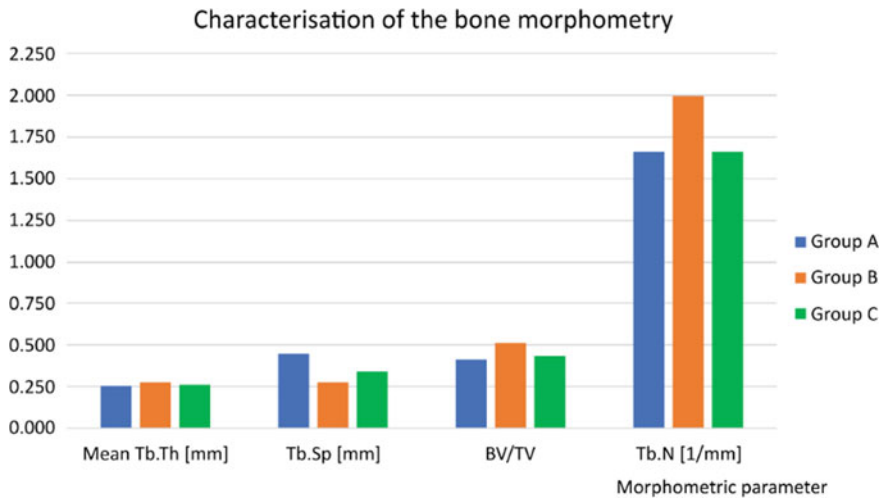
**Table 1** Morphometric parameters of analyzed bone tissue

	Sample	TIGR (%)	Mean Tb.Th (mm)	Tb.Sp (mm)	BV/TV	Tb.N (1/mm)	Conn.D (mm <sup>-3</sup> )	DA (-)
Group A	1	1.6	0.27	1.06	0.20	0.75	2.96	0.38
	2	7.1	0.23	0.25	0.48	2.08	5.17	0.16
	3	4.8	0.28	0.42	0.40	1.42	3.77	0.22
	4	3.6	0.23	0.26	0.46	2.04	3.65	0.20
	5	22.7	0.25	0.24	0.52	2.02	5.16	0.10
	Average	8.0	0.25	0.45	0.41	1.66	4.14	0.21
Group B	1	16.3	0.23	0.19	0.55	2.38	3.95	0.21
	2	0.0	0.43	0.53	0.45	1.04	1.37	0.25
	3	1.5	0.20	0.21	0.48	2.43	4.55	0.19
	4	29.8	0.31	0.17	0.64	2.10	2.99	0.23
	5	28.8	0.21	0.28	0.42	2.03	2.78	0.15
	Average	15.3	0.28	0.28	0.51	2.00	3.13	0.21
Group C	1	0.2	0.27	0.33	0.46	1.66	2.91	0.31
	2	4.7	0.27	0.36	0.42	1.59	2.71	0.11
	3	2.2	0.26	0.31	0.46	1.76	2.75	0.16
	4	3.1	0.20	0.33	0.38	1.86	3.40	0.10
	5	4.0	0.32	0.38	0.45	1.43	2.57	0.15
	Average	2.8	0.26	0.34	0.43	1.66	2.87	0.17

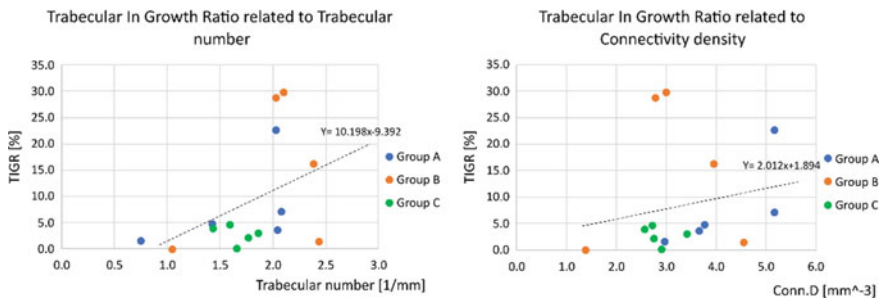
This fact can indicate that the trabeculae have the same thickness within all groups and do not affect the quality of intervertebral fusion.

## 5 Newly Formed Bone Evaluation

It is interesting to look into the relationship between the TIGR value representing the amount of the fused area and morphometric parameters describing the bone properties. There is evident a linear relationship between the TIGR value and Tb.N and Conn.D, respectively. Increasing the fused bone ratio (the TIGR value) also increases the number of trabeculae and their connectivity density. It means that the bone in the LIF area expands as a connected unit. Bone expansion takes place so that the newly formed bone attaches to both vertebrae (in high TIGR values). If the newly formed bone were attached only to one vertebra, the TIGR value would be small (Fig. 4).



**Fig. 3** Graph comparing selected morphological parameters of newly formed bone in the location of LIF area. Individual color bars represent individual groups of samples



**Fig. 4** Graphs depicting the relationship between the ratio of the fused bone (*TIGR* trabecular in growth ratio) and the trabecular number, connectivity density respectively

## 6 Conclusion and Future Work

In this paper, we proposed a methodology to automatically characterize the morphometry of the bone in the fused region after the LIF surgery. The study was elaborated using the samples of porcine vertebrae, where the LIF was conducted using three different types of implants. The main benefit of the proposed methodology is an automatic 3D approach for the evaluation of bone tissue. Automated characterization of fused bone is suitable for accurate comparison of samples where vertebrae are fused with different types of intervertebral implants. The analysis is not affected by operator-induced inaccuracies and is therefore suitable for the inter-laboratory evaluation of osteogenesis bioimplant function in preclinical studies.

In the future, we would like to extend this methodology to the processing of human fused vertebrae samples. Using clinical CT images cannot provide all information mandatory for the analyses described in this paper, but we would utilize the methodology for LIF area extraction and quantify different parameters, such as bone mineral density, bone volume, and detection of fractures or abnormalities in the newly formed bone. The utilization of CT is a standardized diagnostics tool in the pre- and post-surgery diagnosis of LIF. An automated approach for assessing the structure of bone formation between vertebrae may expand the possibilities of diagnosing the success of LIF surgery.

**Acknowledgements** Grant CEITEC VUT-K-22-7761 is realised within the project Quality Internal Grants of BUT (KInG BUT), Reg. No. CZ.02.2.69/0.0/0.0/19\_073/0016948, which is financed from the OP RDE.

## References

1. Fujibayashi, S., et al.: A novel synthetic material for spinal fusion: a prospective clinical trial of porous bioactive titanium metal for lumbar interbody fusion. *European Spine Journal*, 20(9), 1486–1495 (2011), <https://doi.org/https://doi.org/10.1007/s00586-011-1728-3>.
2. Stastny, P., et al.: Structure degradation and strength changes of sintered calcium phosphate bone scaffolds with different phase structures during simulated biodegradation in vitro. *Materials Science and Engineering: C*, 100, 544–553 (2019), <https://doi.org/https://doi.org/10.1016/j.msec.2019.03.027>.
3. Chen, L., et al.: Lumbar interbody fusion with porous biphasic calcium phosphate enhanced by recombinant bone morphogenetic protein-2/silk fibroin sustained-released microsphere: an experimental study on sheep model. *Journal of Materials Science: Materials in Medicine*, 26(3), (2015), <https://doi.org/10.1007/s10856-015-5463-x>.
4. Lo, W.-C., et al.: Understanding the Future Prospects of Synergizing Minimally Invasive Transforaminal Lumbar Interbody Fusion Surgery with Ceramics and Regenerative Cellular Therapies. *International Journal of Molecular Sciences*, 22(7), (2021), <https://doi.org/10.3390/ijms22073638>.
5. Schmidt, C., et al.: Precision and Accuracy of Peripheral Quantitative Computed Tomography (pQCT) in the Mouse Skeleton Compared With Histology and Microcomputed Tomography ( $\mu$ CT). *Journal of Bone and Mineral Research*, 18(8), 1486–1496 (2003), <https://doi.org/10.1359/jbmr.2003.18.8.1486>.
6. He, T., et al.: A comparison of micro-CT and histomorphometry for evaluation of osseointegration of PEO-coated titanium implants in a rat model. *Scientific Reports*, 7(1), (2017), <https://doi.org/10.1038/s41598-017-16465-4>.
7. Lyu, H.-Z., et al.: Correlation between two-dimensional micro-CT and histomorphometry for assessment of the implant osseointegration in rabbit tibia model. *Biomaterials Research*, 25(1), (2021), <https://doi.org/10.1186/s40824-021-00213-x>.
8. Kitchen, D., et al.: Fusion Assessment by MRI in Comparison With CT in Anterior Lumbar Interbody Fusion: A Prospective Study. *Global Spine Journal*, 8(6), 586–592 (2018), <https://doi.org/https://doi.org/10.1177/2192568218757483>.
9. Sethi, A., et al.: Radiographic and CT Evaluation of Recombinant Human Bone Morphogenetic Protein-2-Assisted Spinal Interbody Fusion. *American Journal of Roentgenology*, 197(1), 128–133 (2011), <https://doi.org/https://doi.org/10.2214/AJR.10.5484>.

10. Brans, B., et al.: Assessment of bone graft incorporation by 18 F-fluoride positron-emission tomography/computed tomography in patients with persisting symptoms after posterior lumbar interbody fusion. *EJNMMI Research*, 2(1), (2012), <https://doi.org/10.1186/2191-219X-2-42>.
11. Gadomski, B. et al.: Evaluation of lumbar spinal fusion utilizing recombinant human platelet derived growth factor-B chain homodimer ( rhPDGF-BB ) combined with a bovine collagen/ $\beta$ -tricalcium phosphate (  $\beta$ -TCP ) matrix in an ovine model. *JOR SPINE*, 4(3), (2021), <https://doi.org/10.1002/jsp2.1166>.
12. Tan, G. H., et al.: CT-based classification of long spinal allograft fusion. *European Spine Journal*, 16(11), 1875–1881 (2007), <https://doi.org/https://doi.org/10.1007/s00586-007-0376-0>.
13. Bridwell, K. H., et al.: Anterior Fresh Frozen Structural Allografts in the Thoracic and Lumbar Spine. *Spine*, 20(12), 1410–1418 (1995), <https://doi.org/https://doi.org/10.1097/00007632-199506020-00014>.
14. Laznovsky, J., et al.: Automatic 3D analysis of the ex-vivo porcine lumbar interbody fusion based on X-ray micro computed tomography data. *Computers in Biology and Medicine* 145, (2022), <https://doi.org/10.1016/j.compbiomed.2022.105438>.
15. Gupta, A., et al.: Bone graft substitutes for spine fusion: A brief review. *World Journal of Orthopedics*, 6(6), (2015), <https://doi.org/10.5312/wjo.v6.i6.449>.
16. Jiang, Y., et al.: Application of micro-ct assessment of 3-d bone microstructure in preclinical and clinical studies. *Journal of Bone and Mineral Metabolism*, 23(S1), 122–131 (2005), <https://doi.org/https://doi.org/10.1007/BF03026336>.
17. Boerckel, J. D., et al.: Microcomputed tomography: approaches and applications in bioengineering, 5(6), (2014), <https://doi.org/10.1186/scri534>.
18. Bouxsein, M. L., et al.: Guidelines for assessment of bone microstructure in rodents using micro-computed tomography. *Journal of Bone and Mineral Research*, 25(7), 1468–1486 (2010), <https://doi.org/https://doi.org/10.1002/jbmr.141>.
19. Odgaard, A., Gundersen, H. J. G Quantification of connectivity in cancellous bone, with special emphasis on 3-D reconstructions. *Bone*, 14(2), 173–182 (1993), [https://doi.org/https://doi.org/10.1016/8756-3282\(93\)90245-6](https://doi.org/https://doi.org/10.1016/8756-3282(93)90245-6).
20. Odgaard, A. Three-dimensional methods for quantification of cancellous bone architecture. *Bone*, 20(4), 315–328 (1997), [https://doi.org/https://doi.org/10.1016/S8756-3282\(97\)00007-0](https://doi.org/https://doi.org/10.1016/S8756-3282(97)00007-0).
21. Vojtova, L., et al.: Healing and Angiogenic Properties of Collagen/Chitosan Scaffolds Enriched with Hyperstable FGF2-STAB® Protein: In Vitro, Ex Ovo and In Vivo Comprehensive Evaluation. *Biomedicines*, 9(6), (2021), <https://doi.org/10.3390/biomedicines9060590>.
22. Krticka, M., et al.: Lumbar Interbody Fusion Conducted on a Porcine Model with a Biore-sorbable Ceramic/Biopolymer Hybrid Implant Enriched with Hyperstable Fibroblast Growth Factor 2. *Biomedicines*, 9(7), (2021), <https://doi.org/10.3390/biomedicines9070733>.
23. Domander, R., et al.: BoneJ2 - refactoring established research software. *Wellcome Open Research*, 6, (2021), <https://doi.org/10.12688/wellcomeopenres.16619.2>.
24. Parkinson, I. H., et al.: Variation in segmentation of bone from micro-CT imaging: implications for quantitative morphometric analysis, 31(2), 160–164 (2008), <https://doi.org/https://doi.org/10.1007/BF03178592>.

# Community Detection in Medical Image Datasets: Using Wavelets and Spectral Methods



Roozbeh Yousefzadeh

**Abstract** Medical image datasets may contain a large number of images representing patients with different health conditions. When dealing with raw unlabeled datasets, the large number of samples often makes it hard for experts and non-experts to understand the variety of images present in a dataset. Here, we propose an algorithm to facilitate the automatic identification of communities in medical image datasets. We further demonstrate that such analysis can be insightful in a supervised setting when the images are already labeled. Such insights are useful because health and disease severity can be considered a continuous spectrum. In our approach, we use wavelet decomposition of images in tandem with spectral methods. We show that the eigenvalues of a graph Laplacian can reveal the number of notable communities in an image dataset. Moreover, analyzing the similarities may be used to infer a spectrum representing the severity of the disease (code is available at [https://github.com/roozbeh-yz/community\\_medical\\_images](https://github.com/roozbeh-yz/community_medical_images)).

**Keywords** Unsupervised learning · Medical images · Wavelets · Spectral methods

## 1 Introduction

Analyzing the contents of medical image datasets is not a straightforward task. In practice, it is useful to label images based on health or severity of the disease. Although health and disease can be considered a continuous spectrum, for practical purposes, we usually need to divide that spectrum into specific groups/labels. For example, in the case of analyzing chest X-ray images with respect to the COVID-19 disease, it is useful to define labels such as healthy, mild, severe, and pneumonia. This is not motivated by machine learning, rather by different categories of medical procedures that should follow.

---

R. Yousefzadeh (✉)

Yale Center for Medical Informatics, Yale University, New Haven, CT 06510, USA

e-mail: [roozbeh.yz@gmail.com](mailto:roozbeh.yz@gmail.com)

VA Connecticut Healthcare System, West Haven, CT 06516, USA

**Dealing with unlabeled image datasets.** Annotating and labeling medical images require medical expertise and it is an expensive procedure, prone to mistakes and noisy labels [25]. This makes it sometimes prohibitive to create and analyze large medical image datasets for machine learning and automation. Despite all these difficulties, medical institutions have plenty of raw medical images available, and automating the process of analyzing such datasets and identifying groups of similar images can be beneficial for two reasons.

First, such analysis provides insights about the variety of images present in a raw dataset. For example, if we gather the chest X-ray images of all COVID-19 patients in a given hospital at a given day, it would not be clear how much variety will be present in the gathered data. It would be useful to estimate how many communities of similar images are present in a dataset before having an expert looking over all the images. We show that eigenvalue analysis of a graph Laplacian can provide an estimate of the number of such communities.

Second, automatically detecting groups of similar images can facilitate the labeling process, because the medical expert can then review the groups of images, instead of going through all the images one by one. Here, we show that wavelet decomposition of images in tandem with clustering can facilitate that.

**Dealing with labeled image datasets.** After a medical image dataset is labeled, or when we are given a labeled image dataset, it would be useful to analyze the similarities within each class, and also analyze the cross-class similarities. Mistakes in labeling are not unusual, even by experts, especially when dealing with large datasets. Analyzing the similarities may be able to identify such mistakes. An image that is isolated and dissimilar from other images in a class might actually be a mislabeled image; and even when such images are correctly labeled, it would still be useful to be informed about their existence, and understand the reason behind their dissimilarity to other images of the class. In fact, identifying dissimilar images of the same class are useful for efficient training of models, e.g., triplet mining [11, 24]. Moreover, analyzing the similarities of labeled images may help us automatically infer a disease spectrum representing the severity of disease among patients as we discuss in our results.

**Related work.** There are a few studies that have used community detection methods to detect specific items in images [12, 16]. In those approaches, each community consists of certain pixels inside an image and not a group of similar images inside a dataset. Trivizakis et al. [20] used wavelets to extract features from images, and then, used those features to train a classification model on histology images of colorectal cancer. This shows the effectiveness of wavelets in extracting features. However, their method is not comparable to ours as their focus is on training a classification model on a labeled dataset, not identifying groups of similar images, analyzing in-class and cross-class similarities, and inferring a disease spectrum.

There is a rich literature on community detection algorithms for tabular data and networks [15, 19], but those methods are not readily applicable to image datasets. In the object recognition literature, there are methods that create an embedding for images, but their computational method significantly differs from ours. First, they



are not for medical image datasets, rather for object recognition. Second, they are not concerned with detecting communities of similar images in the datasets. Third, they do not use spectral methods to analyze the abundance of similarities. Fourth, they compare images either by solving expensive optimization problems [21], or by comparing image representations in an inner layer of a trained deep network [3]. This last approach requires a trained model in the first place which can be very expensive.

In previous work, we showed that for object recognition datasets, wavelet decomposition of images can detect similar images in a dataset, the same way that a trained deep learning model does [22]. We also used wavelet decomposition of images to extract independent patterns from image datasets [23].

Recently, Das and Dutta [6] suggested a method to identify images in a medical image dataset most similar to a specific query image. This method is specifically designed for histology images of breast. It uses wavelets to identify specific patches in images, and eventually trains a convolutional neural network and uses the representations learned by the model to identify similar images to the query image. Although this method has similarities to our method, it also has considerable differences. First, it requires training a neural network on images. Second, it is specific to histology images of breast and detection of mitotic cells. Third, it only identifies images similar to a single query image, and does not analyze the similarities in the entire dataset while we perform that task by forming a graph Laplacian for entire datasets and analyzing the eigenvalues.

## 2 Our Method

**Wavelets.** Wavelets are a class of functions and one of the most capable tools to systematically process images and extract features from them. The difficulty of working with images and many signals arises from the spatial complexity of patterns and structures in them. What makes an X-ray image to represent signs of pneumonia cannot be explained by one or a few pixels, rather, it may be explained by the specific patterns that appear in various regions of an image.

Wavelets were developed building on the scientific knowledge of Fourier transform in the context of image and signal processing. Notably, Daubechies [7] showed that wavelets perform better than windowed Fourier transform on visual signals, because wavelets handle the frequencies in a nonlinear way. The family of Daubechies wavelets are one of the most successful types of wavelet transformation, and we use them in this paper. The orthogonality of Daubechies wavelets is particularly useful for feature extraction, because orthogonality in this setting implies the filters are independent and each filter is measuring a specific feature in the image signals. To process images with wavelets, we use the function

$$[\omega, \beta] = \text{wavedec}(x, \Omega, N), \quad (1)$$

which takes as input an image  $x$ , a wavelet basis  $\Omega$ , and level number  $N$ . It returns a vector of real numbers  $\omega$ , representing the wavelet coefficients obtained from convolving  $x$  with  $\Omega$ , and a book keeping matrix  $\beta$  containing the dimensions of wavelet coefficients by level. This operation is reversible, therefore, given  $\omega$ ,  $\beta$ , and  $\Omega$ , we can return to pixel space and reconstruct the image  $x$ , which we denote its operation by

$$x = \text{waverec}(\omega, \beta, \Omega). \quad (2)$$

For a given  $N$ ,  $\beta$  will be constant for all images of the same size.

**Radiomics.** Radiomics refers to an emerging class of computational methods that aim to extract features from medical images that can be useful for clinical decision making and outcome prediction [8, 18]. In certain applications, these methods have been able to extract, from images, features that are not easily detectable by eye, and they have been able to characterize clinically useful phenotypes, e.g., [14]. Computing the radiomic features usually relies on statistical methods and consider the shapes, intensities, textures of items in the images [18]. In a few occasions, more sophisticated methods such as wavelets are used to extract radiomic features, e.g., [1].

**Our Algorithm.** First step is to decompose all images using a wavelet basis  $\Omega$ , and organize them in a matrix  $\mathcal{C}$  with rows corresponding to images and columns corresponding to wavelet coefficients (lines 1–4). About the choice of  $\Omega$ , we have observed empirically that Daubechies-1, 2, and 3, work well in extracting features from images. Higher level wavelets can extract some extra features corresponding to finer details in images, but those finer details are not always useful for analyzing the similarities of images. In fact, when we use higher level wavelets, like Daubechies-5, and extract more features from images, the feature selection part of our algorithm discards those extra features. For each dataset, we recommend a few different wavelet bases to be tested, starting from Daubechies-1. Overall, the choice of  $\Omega$  does not affect our empirical results.

Next, our algorithm selects a subset of wavelet coefficients according to their Laplacian score [9] and using the function `fsulaplacian(·)` (line 5). Laplacian score is a feature selection method based on Laplacian eigenmaps and Locality Preserving Projection [10], specifically designed for unsupervised settings. Wavelet coefficients with scores less than the threshold  $\tau$  will be discarded (line 6). Feature selection with Laplacian score is a standard method and there are standard recommendations for the choice of  $\tau$ . We recommend several values to be tested for  $\tau$  to make sure useful features are not discarded. It is possible to use other feature selection methods as well. In the past, we have used rank-revealing QR factorization [4], but we prefer the Laplacian score because it relates to later steps of our algorithm where we derive the graph Laplacian.

Our algorithm then computes a distance matrix,  $D$ , by applying the function `pdist(·)` on  $\mathcal{C}$  (line 7). `pdist(·)` measures the pairwise distances between the rows of  $\mathcal{C}$  and returns a symmetric square matrix  $D$ . To measure the distances, we use the distance metric  $\mathcal{M}$ . In practice, we have found the *correlation distance* to be an effective metric. Other metrics such as cosine similarity may work as well. We then

convolve the  $D$  with a Gaussian kernel to turn it into an affinity matrix,  $W$  (line 8). In this line,  $std(\cdot)$  returns the standard deviation,  $exp(\cdot)$  is the exponential function, and  $\odot$  is the Hadamard product. The diagonal elements of  $W$  are set to zero.

---

**Algorithm 1** Wavelet Spectral Decomposition for Community Detection (WSDCD): Algorithm for detecting communities of similar images in datasets

---

**Inputs:** Dataset of images  $\mathcal{P}$ , wavelet basis  $\Omega$ , distance metric  $\mathcal{M}$ , feature selection threshold  $\tau_w$ , eigenvalue threshold  $\tau_c$

**Outputs:** Communities in the dataset  $idc$

```

1: Count total number of images in  $\mathcal{P}$  as  $n$ 
2: for  $i = 1$  to  $n$  do
3:    $\mathcal{C}(j, :, i) = \text{wavedec}(\mathcal{P}\{i\}, \Omega)$ 
4: end for
5:  $[idw, scorew] = \text{fsulaplacian}(\mathcal{C})$ 
6:  $\mathcal{C}(:, idw(scorew < \tau_w)) = []$ 
7:  $D = \text{pdist}(\mathcal{C}, \mathcal{M})$ 
8:  $W = \frac{1}{\text{std}(S)} \exp(S \odot S)$ 
9:  $L = \text{glaplacian}(W)$ 
10:  $\lambda = \text{eig}(L)$ 
11: estimate the number of clusters,  $n_c$ , based on the eigen-gaps
12:  $idc = \text{cluster}(\mathcal{C}, \mathcal{M})$ 
13: return  $idc$ 

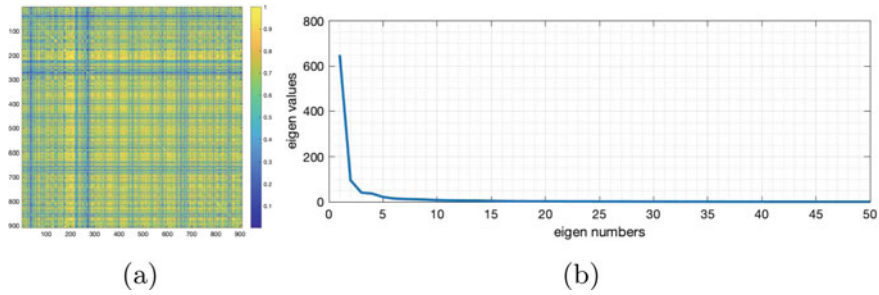
```

---

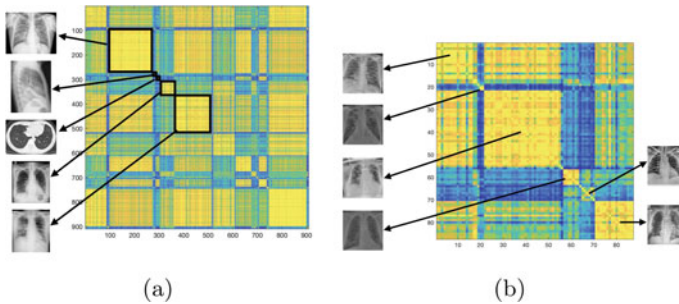
Using the affinity matrix and the function  $\text{glaplacian}(\cdot)$ , we derive the graph Laplacian of the data (line 9). The eigenvalues of the graph Laplacian will let us identify the number of clusters in the data,  $n_c$  (lines 10–11). This is a standard method suggested by von Luxburg [17]. To estimate the number of clusters, it is possible to use alternative methods as well. Finally, we cluster the images into  $n_c$  clusters based on their affinities captured in  $W$  and using a clustering function of choice (line 12). As a result, similar images will appear in each of the clusters and we will be able to provide them to medical experts for further analysis.

### 3 Results

**Dataset on COVID-19 Radiology.** We use the dataset provided by Cohen et al. [5] which contains a mixture of chest X-ray and CT-scan images of patients diagnosed with COVID-19. We proceed with analyzing the dataset by first decomposing the images with Daubechies-3 wavelets. We then measure the cosine similarity of wavelet coefficients of the images. Figure 1a shows the similarity matrix obtained from this analysis. Using the similarity matrix, we then compute its normalized graph Laplacian. Figure 1b shows the eigenvalues of the Laplacian. As we can see, the number of large eigenvalues are not many. In fact, the eigenvalues beyond the



**Fig. 1** Similarity matrix obtained based on wavelet decomposition of all images in the COVID-19 dataset and the distribution of its eigenvalues



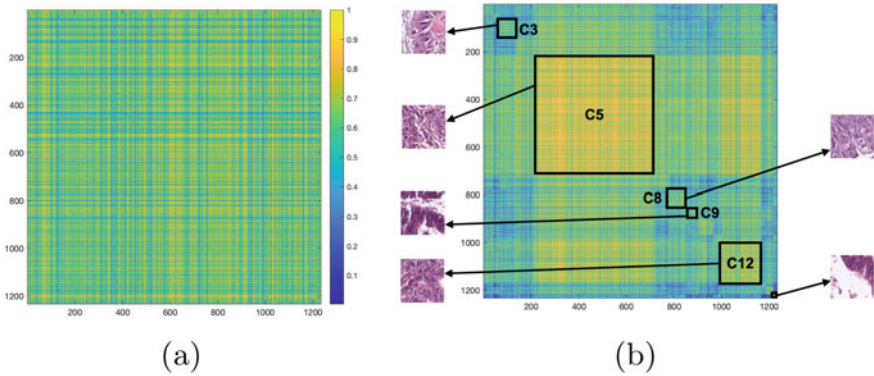
**Fig. 2** **a** Similarity matrix in Fig. 1a after reordering the images based on spectral clustering. Images of different mode appear in different clusters, so does images with different severity of disease. **b** Subset of similarity matrix for images annotated with pneumonia

25th are very close to zero. Based on this, we choose the number of clusters (i.e., image communities) as 25, and proceed with spectral clustering of the images.

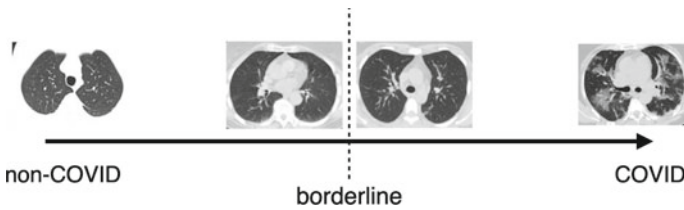
Figure 2a shows the same similarity matrix as in Fig. 1a after re-ordering the rows and columns of the matrix based on the appearance of images in the clusters. Each block along the diagonal of the matrix corresponds to one of the clusters in our image dataset. The off-diagonal blocks reveal the similarity of clusters with each other. Further examination of these clusters reveal that patients with pneumonia appear only in 6 of the 25 clusters, as shown in Fig. 2b. We also see that images of different modality appear in separate clusters.

**Histological images of colorectal cancer (CRC).** Here, we study the colorectal cancer (CRC) histological image dataset [13]. This dataset contains labeled images corresponding to 9 different types of tissue. We use our algorithm to understand the variety of images within the last class of tissues labeled as colorectal adenocarcinoma. Figure 3a shows the resulting similarity matrix. Using the eigenvalues of graph Laplacian, we choose the number of clusters to be 15.

Figure 3b shows the reordered similarity matrix after the clustering and also samples from each of the cluster. Note that all images in all the clusters are considered



**Fig. 3** **a** Similarity matrix for the colorectal adenocarcinoma epithelium class. **b** Reordered similarity matrix based on spectral clustering



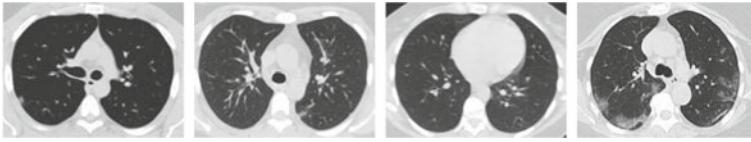
**Fig. 4** Disease spectrum inferred from labeled images. Borderline separates COVID and non-COVID patients. The far right of the spectrum implies high severity of disease, and the far left of the spectrum implies no infection

one malignant type of cancerous tissue. But, there are still different varieties in their patterns. Each cluster appears as a diagonal block in the similarity matrix. By looking at the off-diagonal blocks of the matrix, we can identify which clusters are more similar to each other. For example, note that cluster C3 is more similar to cluster C5 compared to other clusters.

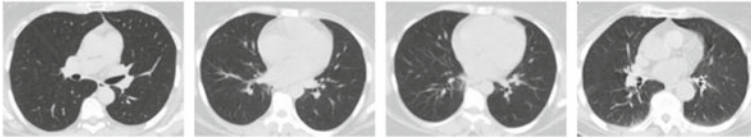
## 4 Inferring the Disease Spectrum

Here, we leverage the similarities and dissimilarities among images to place them on a spectrum representing the severity of disease. The idea is to analyze the similarities of images from two different classes to automatically infer the disease spectrum. Figure 4 shows the disease spectrum we infer for a dataset of SARS-COV-2 CT-Scans [2].

The disease spectrum in Fig. 4 has a borderline in the middle separating CT-scans of patients with COVID from healthy patients. Images of each class that are similar to images of other class will appear near the borderline. Images in one class that have



**Fig. 5** Borderline images: images in the COVID class that are similar to images in the non-COVID class



**Fig. 6** Borderline images: images in the non-COVID class that are most similar to images in the COVID class

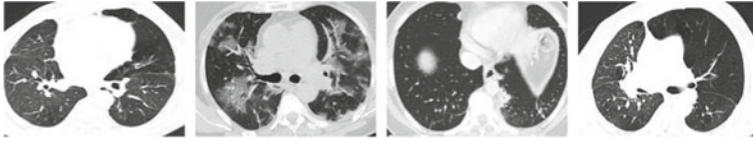
large similarities to images in the other class would be considered in the middle of the disease spectrum, while images that have strong in-class similarities and weak out-class similarities would be placed away from the mid-spectrum, i.e., the borderline. This can be considered an unsupervised approach on labeled images with the aim to extract extra information from them. Labels define which patients have COVID-19, but they do not reveal severity.

To infer a disease spectrum, we investigate images that have considerable similarities to images in the other class. For example, images in the COVID class that are similar to images in the non-COVID class may correspond to patients that are moderately ill. Similarly, images in the non-COVID class that are similar to images in the COVID class may correspond to patients that have vague symptoms of infection. So, we extend our analysis to measure the similarities across classes. Figure 5 shows images in the COVID class that are most similar to images in the non-COVID class, and Fig. 6 shows images in the non-COVID class that are most similar to images in the COVID class. We consider these images to be at the borderline of the spectrum.

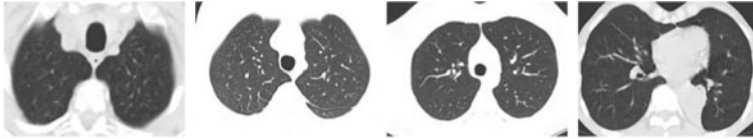
Additionally, Fig. 7 shows images in the COVID class that have the least similarity to images in the non-COVID class. We can consider these images likely to correspond to the infected side of the spectrum. Figure 8 shows images in the non-COVID class that have the least similarity to images in the non-COVID class. We can consider these images likely to correspond to non-infected side of the disease spectrum, far from the borderline.

## 5 Conclusions

We considered a practical setting where a large dataset of medical images is gathered in a medical institution and we need to detect communities of similar images in order



**Fig. 7** COVID images that are most dissimilar to non-COVID images. These may correspond to the infected side of the spectrum, far from the borderline



**Fig. 8** Non-COVID images that are most dissimilar to COVID images. These may correspond to the non-infected side of the spectrum

to proceed with classifying/labeling them. Our algorithm has implications for both unsupervised and supervised learning of medical images. For unsupervised learning, it facilitates the detection of communities of similar images in medical image datasets, improving the expensive process of labeling raw datasets. For supervised learning, our method can help in understanding fine-level similarities within each class and across classes. Such fine-level similarities can be used for training tasks such as triplet mining. Identifying images at the borderline of classes and flagging them for further review by medical experts may reduce the false predictions of deep learning models and make the automated process more reliable. Finally, we showed that analyzing the similarities may be used to infer a disease spectrum.

**Acknowledgements** R. Yousefzadeh thanks Amy Justice and Tamar Taddei for helpful discussions. R. Y. was supported by a fellowship from the Department of Veterans Affairs. The views expressed in this manuscript are those of the author and do not necessarily reflect the position or policy of the United States government.

## References

1. Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications* **5**(1), 1–9 (2014)
2. Angelov, P., Almeida Soares, E.: Explainable-by-design approach for COVID-19 classification via CT-scan. medRxiv (2020)
3. Birodkar, V., Mobahi, H., Bengio, S.: Semantic redundancies in image-classification datasets: The 10% you don't need. arXiv preprint [arXiv:1901.11409](https://arxiv.org/abs/1901.11409) (2019)
4. Chan, T.F.: Rank revealing QR factorizations. *Linear Algebra and its Applications* **88**, 67–82 (1987)
5. Cohen, J.P., Morrison, P., Dao, L.: Covid-19 image data collection. arXiv 2003.11597 (2020). <https://github.com/ieee8023/covid-chestxray-dataset>

6. Das, D.K., Dutta, P.K.: Efficient automated detection of mitotic cells from breast histological images using deep convolution neural network with wavelet decomposed patches. *Computers in Biology and Medicine* **104**, 29–42 (2019)
7. Daubechies, I.: The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory* **36**(5), 961–1005 (1990)
8. Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: Images are more than pictures, they are data. *Radiology* **278**(2), 563–577 (2016)
9. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. *Advances in Neural Information Processing Systems* **18**, 507–514 (2005)
10. He, X., Niyogi, P.: Locality preserving projections. *Advances in Neural Information Processing Systems* **16**(16), 153–160 (2004)
11. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017)
12. Javed, S., Mahmood, A., Fraz, M.M., Koohbanani, N.A., Benes, K., Tsang, Y.W., Hewitt, K., Epstein, D., Snead, D., Rajpoot, N.: Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Medical Image Analysis* **63**, 101,696 (2020)
13. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine* **16**(1), e1002,730 (2019)
14. Lafata, K.J., Zhou, Z., Liu, J.G., Hong, J., Kelsey, C.R., Yin, F.F.: An exploratory radiomics approach to quantifying pulmonary function in CT images. *Scientific Reports* **9**(1), 1–9 (2019)
15. Li, Y., He, K., Kloster, K., Bindel, D., Hopcroft, J.: Local spectral clustering for overlapping community detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **12**(2), 1–27 (2018)
16. Linares, O.A., Botelho, G.M., Rodrigues, F.A., Neto, J.B.: Segmentation of large images based on super-pixels and community detection in graphs. *IET Image Processing* **11**(12), 1219–1228 (2017)
17. von Luxburg, U.: A tutorial on spectral clustering. *Statistics and Computing* **17**(4), 395–416 (2007)
18. Rizzo, S., Botta, F., Raimondi, S., Origgi, D., Fanciullo, C., Morganti, A.G., Bellomi, M.: Radiomics: the facts and the challenges of image analysis. *European Radiology Experimental* **2**(1), 1–8 (2018)
19. Shi, P., He, K., Bindel, D., Hopcroft, J.E.: Locally-biased spectral approximation for community detection. *Knowledge-Based Systems* **164**, 459–472 (2019)
20. Trivizakis, E., Ioannidis, G.S., Souglakos, I., Karantanas, A.H., Tzardi, M., Marias, K.: A neural pathomics framework for classifying colorectal cancer histopathology images based on wavelet multi-scale texture analysis. *Scientific Reports* **11**(1), 1–10 (2021)
21. Vo, H.V., Bach, F., Cho, M., Han, K., LeCun, Y., Pérez, P., Ponce, J.: Unsupervised image matching and object discovery as optimization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8287–8296 (2019)
22. Yousefzadeh, R.: Using wavelets to analyze similarities in image-classification datasets. *arXiv preprint arXiv:2002.10257* (2020)
23. Yousefzadeh, R., Huang, F.: Using wavelets and spectral methods to study patterns in image-classification datasets. *arXiv preprint arXiv:2006.09879* (2020)
24. Yuan, Y., Chen, W., Yang, Y., Wang, Z.: In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 354–355 (2020)
25. Zhang, Y., Wei, Y., Wu, Q., Zhao, P., Niu, S., Huang, J., Tan, M.: Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Transactions on Image Processing* **29**, 7834–7844 (2020)



# Non-pooling Network for Medical Image Segmentation



Wei hu Song, Heng Yu, and Jianhua Wu

**Abstract** Existing studies tend to focus on model modifications and integration with higher accuracy, which improve performance but also carry huge computational costs, resulting in longer detection times. In medical imaging, the use of time is extremely sensitive. And at present most of the semantic segmentation models have encoder-decoder structure or double branch structure. Their several times of the pooling use with high-level semantic information extraction operation cause information loss although there is a reverse pooling or other similar action to restore information loss of pooling operation. In addition, we notice that visual attention mechanism has superior performance on a variety of tasks. Given this, this paper proposes non-pooling network (NPNet), non-pooling commendably reduces the loss of information and attention enhancement module (AEM) effectively increases the weight of useful information. The method greatly reduces the number of parameters and computation costs by the shallow neural network structure. We evaluate the semantic segmentation model of our NPNet on three benchmark datasets comparing with multiple current state-of-the-art (SOTA) models, and the implementation results show that our NPNet achieves SOTA performance, with an excellent balance between accuracy and speed.

**Keywords** Medical image segmentation · Non-pooling · Deep learning · Attention enhancement module

---

W. Song (✉)

Beihang University, 37 Xueyuan Road, Haidian District, Beijing, China

e-mail: [weihusong@buaa.edu.cn](mailto:weihusong@buaa.edu.cn)

H. Yu

Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

J. Wu

NanKai University, No. 38, Tongyan Road, Haihe Education Park, Haidian District, Tianjing, China

## 1 Introduction

Medical image segmentation can promote the research and development of medical field. It can help doctors analyse and take action using image features. The accuracy and speed of image segmentation is critical and existing research is carried out from these two aspects. However, the relationship between accuracy and speed in most models has not reached a relative balance. FCN [1], as the first semantic segmentation model, undoubtedly attracted great attention. It changed the last full connection layer of the classification network into convolution to achieve remarkable performance on semantic segmentation. U-Net [2], the first segmentation network proposed for medical images, is a typical encoder-decoder structure. In U-Net, skip connections are used to effectively integrate shallow spatial information and deep semantic information, thus making up for the loss of feature information caused by multiple pooling operations in the encoder stage. Subsequently, a series of improved models based on U-Net show up. More complex feature extraction modules are used to extract as much feature information as possible from each level of the segmentation network to weaken the influence of pooling operations on information loss. Although such models improve certain performance, However, more redundant information, even error information, was introduced, and the model size and computation cost also increased significantly, creating a certain burden. SegNet [3] uses two lassification networks as encoder and decoder respectively and proposes to use max pool index to do up-sampling to better restore the impact of pooling. PSPNet [4] and DeePLab series [5–8] both use image classification networks as the backbone. The former proves the effectiveness of extracting multi-size feature information by pyramid pooling module for the first time, while the latter proposes to use atrous convolution to obtain feature information of larger receptive field, using atrous spatial pyramid pooling (ASPP) to obtain rich feature information of multiple dimensions. In addition, the attention mechanism introduced from NLP to computer vision has also shown its dazzling brilliance, among which SENet [9] is undoubtedly the most important representative, its excellent performance won the last imagenet champion. Some other semantically segmented networks use dual branching structures [10] to acquire semantic information [10], spatial information [10], and cascade structures [11] respectively. Above all, most of the semantic segmentation models have encoder-decoder structure, use image classification network as the backbone, and use the structure of the double branch or cascade structure. Pooling operation used in these semantic segmentation models leads to information loss. And the complex structure will also cause the burden of model and calculation. In addition, dilated convolution, multi-dimensional feature extraction, and attention mechanisms are proved to be effective. Therefore, we elaborately designed a simple and novelty non-pooling network, which solves the information loss caused by the pooling operation and uses improved ASPP and a new plug-and-play attention mechanism module. Our contribution is in four aspects: (1) We propose a new plug-and-play attention mechanism module, which has better performance than the attention module in SENet. (2) We propose an improved ASPP module and have better performance. (3) For the

first time, non-pooling semantic segmentation model with only 1/50 of the number of model parameters and computation costs of U-Net is proposed. (4) Our method surpasses other state-of-the-art performance on three medical image segmentation datasets.

## 2 Methodology

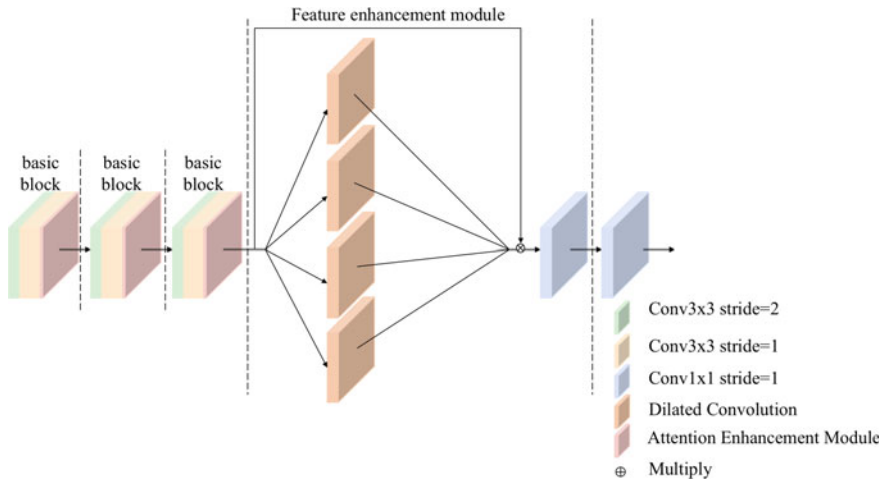
In this paper, we propose NPNet, a novel lightweight semantic segmentation model for medical images. The network structure is described in Fig. 1. It is mainly composed of the basic block, attention enhancement module shown in Fig. 1, and feature enhancement module. In this model, all the convolution operations are  $3 \times 3$  convolution kernel, followed by batch normalization(bn) and ReLU. There are three basic blocks at the beginning of the network, and attention enhancement module is added after each block, and a feature enhancement module is implemented in the middle of the network. At the back of the network,  $1 \times 1$  convolution is used to output according to the classification number and bi-linear interpolation is used to restore the original input size. In this section, we will talk about these components in detail.

### 2.1 Basic Block

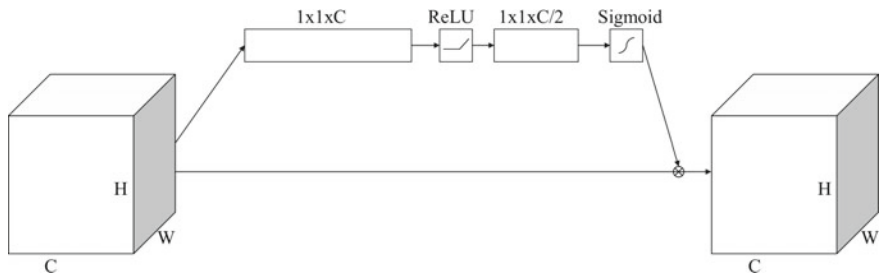
The basic block first uses  $3 \times 3$  convolution operation with stride equal to 2, which is equivalent to the function of reducing image size and computation achieved by the pooling operation of stride equal to 2. Moreover, the convolution operation can also effectively deal with the loss of feature information in pooling operation. Then, two  $3 \times 3$  convolution operations with stride equal to 1 are used to fully extract the feature information of this layer, and also to obtain more abundant and useful feature information which is conducive to the transmission of the information of the next layer. The design of this basic block is to reduce the information loss and realize the effective extraction of spatial feature information of images with different sizes.

### 2.2 Attention Enhancement Module

Attention enhancement module is an attention module integrated with  $1 \times 1$  convolution. First, the input image is transformed into a 1-dimensional matrix by adaptive average pooling, and then the input dimension is transformed into the input dimension divided by parameter reduction using  $1 \times 1$  convolution. Like SENet, we also use



(a) NPNet



(b) attention enhancement module

Fig. 1 Proposed NPNet architecture

linear activation function ReLU and nonlinear activation function Sigmoid successively. The difference is that we use  $1 \times 1$  convolution to replace the full connection layer.  $1 \times 1$  convolution can effectively improve the nonlinear characteristics and information interaction across channels, better extracting useful information in feature information. Finally, the nonlinear activation function Sigmoid is used on output and the result is weighted by multiplying the original input to achieve channel adaptive weighting. This module is used after each basic block to further strengthen the weight of useful information between different channels, thus providing better characteristic information for the following steps.

### 2.3 Feature Enhancement Module

The feature enhancement module is used after three basic blocks, and the image size currently is  $1/8$  of the input size. This module is composed of four dilated convolutions with the ratio of 1, 5, 15, 20, and two  $1 \times 1$  convolutions. The atrous convolution can obtain feature information from the large receptive field without increasing the number of parameters, obtaining richer semantic information. The input image is first output through four dilated convolutions according to  $1/2$  of the number of output channels. Then the four outputs are superimposed through concatenation, and the number of channels is 2 times the number of output channels.  $1 \times 1$  convolution is used to achieve dimension reduction, that is, the normal number of output channels is obtained. At this point, the residual structure is introduced to concatenate the result with the original input information to realize feature reuse. Finally,  $1 \times 1$  convolution is used for further dimension reduction.

## 3 Experiments and Results

### 3.1 Datasets

**Lung Segmentation.** Lung CT image segmentation is an important and initial step in lung CT image analysis. This dataset comes from the Kaggle contest, Finding and Measuring Lungs in CT data1 (Luna for short). It consists of 267 2D images and is randomly split into train set (80%) and test set (20%). Also, we use the original image equally.

**Skin Lesion Segmentation.** Computer-aided automatic diagnosis of Skin cancer is an inevitable trend, and Skin lesions segmentation as the first step is urgent. The data set is from MICCAI 2018 Workshop—ISIC2018: Skin Lesion Analysis Towards Melanoma Detection [12, 13] (Skin for short). It contains 2594 images and is randomly split into train set (80%) and test set (20%). For better model training and result display, we resize all the original images to  $224 \times 224$ .

**Polyp Segmentation.** Accurate detection of colon polyps is of great significance for the prevention of colon cancer. CVC-ClinicDB [14] (CVC for short) includes 612 colon polyp images. We use the original size  $384 \times 288$  of image and split it into train set (80%) and test set (20%).

### 3.2 Experimental Settings

For three benchmarks and multiple segmentation models, we set consistent training parameters. We set epochs as 100 in the three data sets. We use a learningrate (LR)

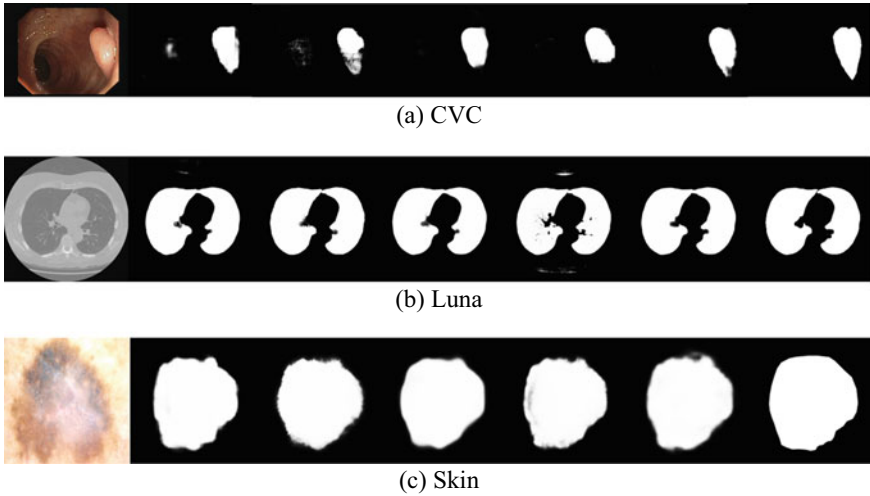
equal to  $1e-3$  for Luna and Skin task and  $1e-4$  for CVC task. In addition, we use batch size equal to 2 for Luna and CVC task, and 4 for the Skin task. Cross entropy loss and Adam are used as loss function and optimizer, respectively. All experiments run on the NVIDIA GeForce RTX 2080Ti GPU with 12 GB. Intersection over Union (IOU), dice coefficient (Dice) and multiply-accumulate operations (MACs) are selected as the evaluation metrics in this paper. We used these evaluation metrics for all datasets.

### 3.3 *Experimental Results*

In this section, we presented qualitative results on three data sets and compared with other SOTA semantic segmentation networks to prove the superior performance of our NPNet. We set up the same parameters for the same data set in different network models, and all models were trained from scratch. Since U-Net is still the baseline of many networks, we also introduce several SOTA models based on U-Net for comparison. Table 2 shows that our model is superior to other SOTA models in terms of performance and is 1/50 of U-Net in terms of model size and computation costs. In all the figures demonstrating the qualitative results in Fig. 2, the sequence are origin image, FCN8s, SegNet, PSPNet, U-Net, NPNet, mask, respectively. It can be found from these figures that the model proposed can effectively reduce the loss of information with our non-pooling module, so as to retain more details and achieve better performance. Moreover, the model size and computation cost of this paper are only 1/100 of U-Net++. Ablation Studies. The attention mechanism module in SENet plays an important role, and many models insert this template into their models to achieve better performance. Therefore, we conducted an experimental comparison on three datasets of our proposed attention enhancement module. The difference between the model size and the computation costs of these three models can be negligible, the experimental results in Table 1 prove that our attention enhancement module is better than SENet as a plug-and-play attention module.

## 4 Conclusion

In this work, we propose a novel semantic segmentation network with non-pooling operation for the first time, which can effectively alleviate the problem of information loss and difficult recovery caused by the pooling operation. Our proposed network also gets rid of common encoding and decoding structures. In addition, we also proposed an attention module to enhance feature information, which can be easily inserted into other network models with fewer parameters, Experiment results on three datasets show that our model can surpass state-of-art counterparts with lightweight parameters and MACs.



**Fig. 2** Qualitative comparison of different segmentation results

**Table 1** Comparison on CVC, Skin and Luna with seven models

Dataset	Methods	IOU	Dice	Params (M)	MACs (G)
	FCN8s [1]	0.6149	0.7249	14.72	33.89
	SegNet [3]	0.7146	0.7933	29.44	67.67
	PSPNet [4]	0.7159	0.8045	17.5	133.23
	U-Net [2]	0.7439	0.8229	34.53	110.46
	Attention U-Net [15]	0.7334	0.8153	34.87	112.27
	U-Net++ [16]	0.7632	0.8356	36.63	233.88
	NPNet	0.7766	0.8397	0.71	2.17
Skin	FCN8s [1]	0.7828	0.8511	14.72	61.50
	SegNet [3]	0.7897	0.8558	29.44	30.71
	PSPNet [4]	0.8052	0.8708	17.5	60.45
	U-Net [2]	0.8086	0.8691	34.53	50.12
	Attention U-Net [15]	0.8028	0.8691	34.87	50.94
	U-Net++ [16]	0.7901	0.8588	36.63	106.11
	NPNet	0.8170	0.8757	0.71	0.99
Luna	FCN8s [1]	0.9741	0.9802	14.72	80.32
	SegNet [3]	0.9688	0.9789	29.44	160.41
	PSPNet [4]	0.9732	0.9823	17.5	315.87
	U-Net [2]	0.9749	0.9821	34.53	261.64
	Attention U-Net [15]	0.9698	0.9794	34.87	266.11
	U-Net++ [16]	0.9746	0.9831	36.63	554.37
	NPNet	0.9785	0.9832	0.71	5.15

**Table 2** Evaluation of proposed attention enhancement module

Dataset	Methods	Attention	IOU	Dice
CVC	NPNet	No	0.7439	0.8157
	NPNet	SENet	0.7448	0.8186
	NPNet	AEM	<b>0.7766</b>	<b>0.8397</b>
Luna	NPNet	No	0.9758	0.9807
	NPNet	SENet	0.9772	0.9820
	NPNet	AEM	<b>0.9785</b>	<b>0.9832</b>
Skin	NPNet	No	0.8131	0.8742
	NPNet	SENet	0.8091	0.8709
	NPNet	AEM	<b>0.8165</b>	<b>0.8766</b>

## References

1. Cheng HK, Chung J, Tai YW, Tang CK (2020) Cascadepsp: Toward class agnostic and very high-resolution segmentation via global and local refinement. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
2. Huang, W., Li, Y., Zhang, K., Hou, X., Xu, J., Su, R. and Xu, H., 2021. An Efficient Multi-Scale Focusing Attention Network for Person Re-Identification. *Applied Sciences*, 11(5), p.2010.
3. Shelhamer E, Long J, Darrell T (2017) Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39:640–651.
4. Tschandl P, Rosendahl C, Kittler H (2018) The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* 5(1):1–9.
5. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39:2481–2495.
6. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 6230–6239.
7. Chen LC, Papandreou G, Kokkinos I, Murphy KP, Yuille AL (2015) Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR abs/1412.7062*.
8. Chen LC, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. *ArXiv abs/1706.05587*.
9. Han W, Zhang Z, Zhang Y, Yu J, Chiu CC, Qin J, Gulati A, Pang R, Wu Y (2020) Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. *arXiv preprint arXiv:200503191*.
10. Jie H, Li S, Gang S, Albanie S (2017) Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP(99).
11. Chen LC, Papandreou G, Kokkinos I, Murphy KP, Yuille AL (2018) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40:834–848.
12. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. *ArXiv abs/1802.02611*.
13. Codella N, Rotemberg V, Tschandl P, Celebi ME, Dusza S, Gutman D, Helba B, Kalloo A, Liopyris K, Marchetti M, et al. (2019) Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:190203368*.
14. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: *MICCAI*.



15. Oktay O, Schlemper J, Folgoc LL, Lee MJ, Heinrich MP, Misawa K, Mori K, McDonagh SG, Hammerla NY, Kainz B, Glocker B, Rueckert D (2018) Attention u-net: Learning where to look for the pancreas. ArXiv abs/1804.03999.
16. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2018) Unet++: A nested u-net architecture for medical image segmentation. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, S 11045:3–11.

# Lung CT Analysis Using 3D Disparity-Regularised Block Matching for Stereotactic Ablative Body Radiotherapy



Durai Arun Pannir Selvam, David I. Laurenson, William H. Nailon,  
and Duncan B. McLaren

**Abstract** Identification and quantification of pulmonary fibrosis and other anomalies are challenging problems in Radiotherapy. This paper introduces a block matching technique that characterises the voxel displacements as a geometrical relationship between lung CT scans. The proposed block matching technique uses two-pass distance and orientation-based regularisation to restrict unnatural and unrealistic tissue deformations. Also, this technique uses both texture maps and voxel intensities as block matching criteria. This yields a displacement vector field whose predicted motion vectors are closer to the actual displacements evaluated at every slice location using image quality-based performance metrics-namely structural similarity index (0.9959), mean squared error (0.0029), and peak signal-to-noise ratio (46.6). Thus providing a quantitative approach to the clinicians aiding in identifying and quantifying the clinically significant geometrical changes, eventually characterising the tumour degradation or pulmonary fibrosis in terms of volumetric and shape changes.

**Keywords** Radiotherapy · Block matching · Motion vector regularisation · Displacement vector field

---

Research supported by IDCom, The School of Engineering, The University of Edinburgh and NHS Scotland.

---

D. A. Pannir Selvam (✉) · D. I. Laurenson · W. H. Nailon  
Institute of Digital Communication, University of Edinburgh, Edinburgh, UK  
e-mail: [da.pannir-selvam@sms.ed.ac.uk](mailto:da.pannir-selvam@sms.ed.ac.uk)

D. I. Laurenson  
e-mail: [dave.laurenson@ed.ac.uk](mailto:dave.laurenson@ed.ac.uk)

W. H. Nailon · D. B. McLaren  
Edinburgh Cancer Centre, Western General Hospital, Edinburgh, UK

## 1 Introduction

The role of image processing techniques in Radiotherapy has become so important that it is the fundamental framework for the Image-guided Radiotherapy (IGRT) [2]. In particular, deformable image registration has been used to improve Image-guided Adaptive Radiotherapy (ART) [11]. Among several negative side effects [1] of ART, the toxicity in patients caused by therapeutic radiation is one of the key limitations, and is therefore a focus of current research [3]. Geometric uncertainties (anatomical changes) in patients is the one of the reasons for the toxicity of organs nearby a cancerous tissue i.e. organs-at-risk (OAR) [5] which can be seen in the scans acquired during the inter-fraction and intra-fraction time period. Therefore, by having a method that characterises these anatomical changes using the geometrical relationship between any two scans will aid in improving the benefits of ART. In that perspective, image registration is used to characterise the geometrical changes and estimate the geometric transform between two scans. This research is a part of the study that assesses the ability to detect pre-treatment tumour cell free DNA (cfDNA) in peripheral blood of patients with early stage lung cancer receiving Stereotactic Ablative Body Radiotherapy (SABR) and the impact of SABR radiotherapy on tumour cfDNA, cardiac cfDNA and lung cfDNA during radical radiotherapy for Non-Small Cell Lung Cancer (NSCLC). In this study, two follow-up lung CT scans are acquired after definite time intervals post therapeutic radiation. Using those follow-up scans, the geometrical changes will be characterised using the image registration techniques to assist the clinical experts to identify the onset of pulmonary fibrosis and other clinically-significant changes. The disparity-regularised block matching based non-rigid registration was able to geometrically characterise the organ deformations occurring in the scans acquired at different time period by restraining unnatural deformations during the motion vector estimation with the two-pass based distance and orientation regularisation [13]. In that sense, the disparity-regularised block matching is modified to identify clinically significant sites in the lung CT images. The proposed modification uses the texture information as the block matching criteria. Using texture information as a part of the block matching has been tried before [4, 16], however, very few investigations [10, 14] have been conducted in terms of estimating a three-dimensional motion vector field for CT scans that has large area of homogeneous and coarse texture like lung CT. In place of the traditional approach of using texture as features in block matching cost function [9, 12], an alternative approach of using them as maps [7] is proposed for the disparity-regularised block matching. This modified disparity-regularised block matching based motion estimation is able to detect movements in the lung CT, consequently, characterising the clinically significant sites and their geometrical changes, with less computational complexity. Therefore, identifying the regions where there are changes in the current follow-up scan by comparing it with the previous follow-up scan is vital. This comparison helps clinicians to assess the disease progression or detect abnormalities that might have appeared within the time period of the scans.

## 2 Methodology

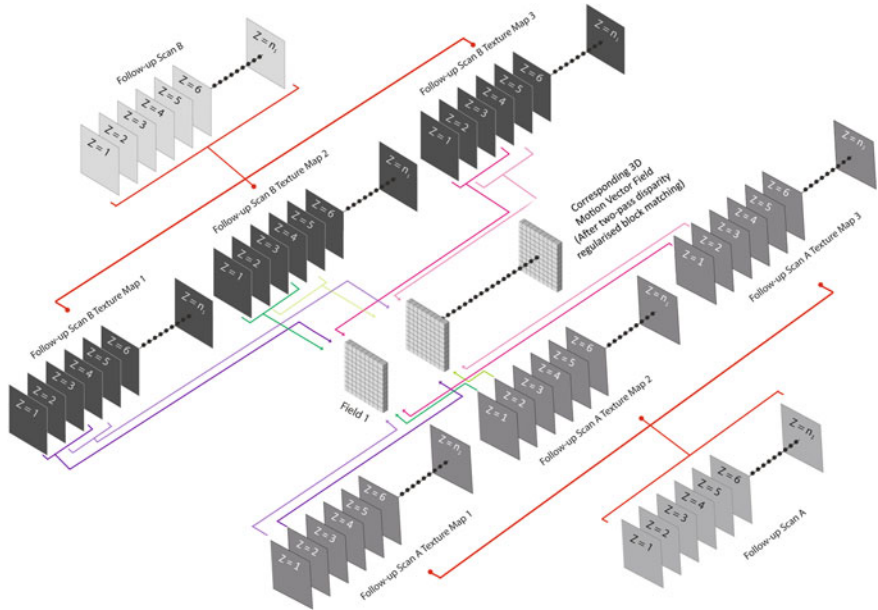
### 2.1 *Materials and Data Preprocessing*

In the cfLungDNA dataset used in this study, there are follow-up CT scans for each patient that were acquired on the 4th and 12th month after the radiation therapy. The follow-up scan  $B$  was acquired on the 4th month post radiation treatment and the follow-up scan  $A$  was acquired on the 12th month post radiation treatment. The voxel size of 4th CT scan is  $0.7559 \text{ mm} \times 0.7559 \text{ mm} \times 2 \text{ mm}$  and the aspect ratio is  $512 \times 512$ . The voxel size of 12th CT scan is  $0.8926 \text{ mm} \times 0.8926 \text{ mm} \times 1 \text{ mm}$  and the aspect ratio is  $512 \times 512$ . From the voxel size and aspect ratio of both scans, it can be observed that the array size and the physical co-ordinates of scans will differ accordingly. Therefore, the follow-up scans  $A$  and  $B$  were rigidly registered to share the same voxel co-ordinate system. Thus rigid registration becoming a pre-requisite for the proposed technique.

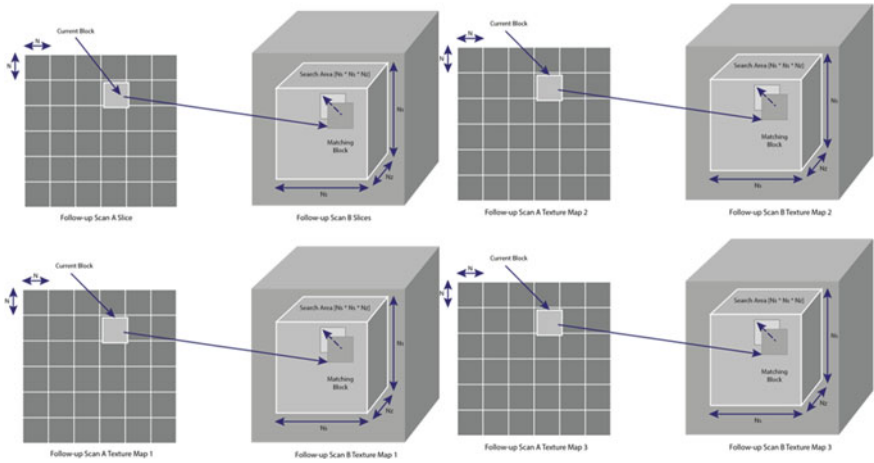
### 2.2 *Lung CT Analysis Using Extended DBLM*

The disparity-regularised block matching technique (DBLM) as laid out in [13] was able to propagate the contours from a planning scan to the on-the-day-of-treatment scan using the geometrical relationship between the scans, provided they were from the same modality. Also, the displacement vector field (DVF) estimation by DBLM was compared with b-spline-a parametric, demons-a non-parametric, and pyramidal block matching (BLMP) which were chosen from clinical evaluation studies [8, 15]. However, due to larger coarser regions in this CT scan, it is evident that the DBLM cannot be used in the same form. Because, the image similarity metric (ISM) i.e. mean absolute error (MAE) calculated between the voxel intensities in the DBLM is very sensitive to the intensity variance (texture). Therefore, the proposed modifications for the DBLM to handle the large coarser regions is to use texture information as a part of the block matching criteria (disparity function) along with the MAE of the voxel intensities, hence the name, Extended DBLM. Instead of using the traditional gray-level co-occurrence matrix (GLCM) as the texture features in the DBLM, texture maps were used in the Extended DBLM. The texture filters are applied on the scan slices and the filter output from these filters are the texture maps. The texture filters chosen for this technique are entropy filter, range filter and standard deviation filter. These three texture filters are chosen because of its common usage in the texture based image analysis [6]. The filter outputs from these texture filters are labelled as Texture Map 1, 2 and 3 in Figs. 1 and 2 to emphasise that any texture filter could be used here and the suitability of other texture filters is yet to be investigated.

The entropy filter calculates the entropy for every  $k$ th slice in the volume and generates the entropy texture map i.e. Texture Map 1. These slice-wise texture maps are then stacked together as the three-dimensional entropy map  $I_x^E$ . Using this range



**Fig. 1** The displacement vector field estimation process using extended disparity-regularised block matching



**Fig. 2** The exhaustive search based block matching used on texture maps in extended disparity-regularised block matching

filter on the slices of the scan volume, it generates slice-wise maps which are stacked as three-dimensional coarseness map  $I_x^R$ . Similar to the other texture filters, the standard deviation filter generates slice-wise maps and stacks them as three-dimensional variability map  $I_x^S$ . Finally, all the texture maps were normalised so that the hyper-parameters of the DBLM could be reused in the Extended version. Then using the 3D normalised texture maps along with the scan volumes, the disparity regularised block matching is performed to estimate the exclusive 3D displacement vector fields for every slice location as illustrated in Fig. 1.

As per the cfLungDNA dataset, two follow-up scan 3D volumetric arrays say  $I_{fA}$  and  $I_{fB}$  per patient are considered for the establishment of a geometrical relationship between them using the Extended DBLM. Therefore, upon  $I_{fA}$  and  $I_{fB}$ , the said texture filters are applied to obtain the three-dimensional entropy maps ( $I_{fA}^E$  and  $I_{fB}^E$ ), three-dimensional coarseness maps ( $I_{fA}^R$  and  $I_{fB}^R$ ), and three-dimensional degree of variability maps ( $I_{fA}^S$  and  $I_{fB}^S$ ) as inputs to the block matching criteria i.e. the disparity function. Similar to the DBLM, the search space for a slice in follow-up scan  $I_{fA}$  slice is chosen from several slices in follow-up scan  $I_{fB}$ . Figure 2, illustrates all the texture map based search spaces and the voxel intensity based search space required for motion vector calculation for one block in scan which can be compared with the search space of DBLM [13]. In the Extended DBLM, a  $b_{fA}[m, n]$  block of size  $N \times N$  belonging to a slice in follow-up scan  $I_{fA}$  is scanned over several  $b_{fB}[m, n, z]$  blocks of size  $N \times N$  in the search area of size  $N_s \times N_s \times N_z$  belonging to the corresponding slices of follow-up scan  $I_{fB}$ , where  $N_s = N + p$ ,  $p$  is the search parameter and  $N_z$  is the number of slices in follow-up scan  $I_{fB}$  per group. Similarly, the blocks from the texture maps belonging to the slice of follow-up scan  $I_{fA}$  i.e.,  $b_{fA}^E[m, n]$ ,  $b_{fA}^R[m, n]$ , and  $b_{fA}^S[m, n]$  belonging to  $I_{fA}^E$ ,  $I_{fA}^R$ , and  $I_{fA}^S$  are scanned over several  $b_{fB}^E[m, n, z]$ ,  $b_{fB}^R[m, n, z]$ , and  $b_{fB}^S[m, n, z]$  belonging to the texture maps of follow-up scan  $I_{fB}$ , say,  $I_{fB}^E$ ,  $I_{fB}^R$ , and  $I_{fB}^S$  respectively. The MAE of all the blocks are aggregated and used as the ISM for all corresponding locations (*Candidate Displacements*) in a search area giving a disparity cost matrix of size  $2p + 1 \times 2p + 1 \times N_z$  which are expressed in Eqs. 1–5,

$$ISM_v(i, j, k) = ||b_{fA}[m, n] - b_{fB}[m, n, z]|| \quad (1)$$

$$ISM_E(i, j, k) = ||b_{fA}^E[m, n] - b_{fB}^E[m, n, z]|| \quad (2)$$

$$ISM_R(i, j, k) = ||b_{fA}^R[m, n] - b_{fB}^R[m, n, z]|| \quad (3)$$

$$ISM_S(i, j, k) = ||b_{fA}^S[m, n] - b_{fB}^S[m, n, z]|| \quad (4)$$

$$ISM(i, j, k) = [ISM_v + ISM_E + ISM_R + ISM_S]/4, \quad (5)$$

where,  $b_{fA}[m, n] \in I_{fA}[N \times N]$ ,  $b_{fB}[m, n, z] \in I_{fB}[N_s \times N_s \times N_z]$ ,  $b_{fA}^E \in I_{fA}^E$ ,  $b_{fB}^E \in I_{fB}^E$ ,  $b_{fA}^R \in I_{fA}^R$ ,  $b_{fB}^R \in I_{fB}^R$ ,  $b_{fA}^S \in I_{fA}^S$ ,  $b_{fB}^S \in I_{fB}^S$ ,  $m$  is a block's rows,  $n$  is

a block's columns,  $I_{fA}$  is a follow-up scan 3D array,  $I_{fB}$  is another follow-up scan 3D array,  $I_x^E$  is a 3D Entropy (Randomness) Texture Map,  $I_x^R$  is a 3D Coarseness Texture Map, and  $I_x^S$  is a 3D Degree of Variability Texture Map. With these equations, the motion vectors are estimated for the whole volume, making the collection of all the motion vectors as the displacement vector field whose magnitude is used for visualisation.

### 2.3 Parameters of Extended DBLM and Its Performance Evaluation

Parameters selected for the proposed disparity-regularised block matching algorithm are block size  $N = 5$  pixels, search space parameter  $p = 5$  pixels, and the step-size for the search is 1 pixel. The number of slices per group  $N_z$  chosen for the transformation is 5. The DVF estimated by the Extended DBLM will be evaluated by image quality based performance metrics namely—structural similarity index metric (SSIM), normalised mean squared error (NMSE) and peak signal-to-noise ratio (PSNR). Since the DVF estimated at every slice location of the volumetric stack, the performance metrics are calculated between slices of the ground-truth follow-up scan  $A$  with the slices of the geometrically warped follow-up scan  $B$ , i.e., the slices of the estimated follow-up scan  $A$ . These metrics are defined in the following equations, where  $I_M$  and  $\hat{I}_M$  represents each slices of the ground-truth 3D image and the estimated 3D image.

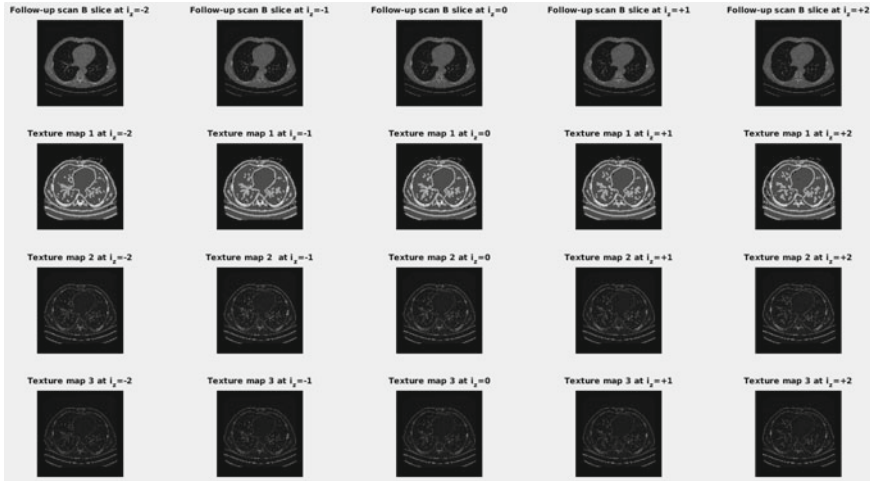
$$MSE = \frac{1}{N_v} \sum_{n=x,y,z}^{N_v} |I_M(x, y, z) - \hat{I}_M(x, y, z)|^2 \quad (6)$$

$$NMSE = \frac{MSE}{\sqrt{E_{I_M}} \times \sqrt{E_{\hat{I}_M}}} \quad (7)$$

$$PSNR = 10 \times \log_{10} \left( \frac{MAX^2}{MSE} \right) \quad (8)$$

## 3 Results and Discussion

The Extended DBLM obtains the texture maps using the texture filters, such as entropy maps  $I_{fA}^E$  and  $I_{fB}^E$ , coarseness maps  $I_{fA}^R$  and  $I_{fB}^R$ , and degree of variability maps  $I_{fA}^S$  and  $I_{fB}^S$ , from 3D volumetric arrays  $I_{fA}$  and  $I_{fB}$  representing the follow-up scans  $A$  and  $B$ . With  $N_z$  as 5 here, for a slice location  $n$  in the follow-up scan  $A$ , the slice locations of the corresponding search group in follow-up scan  $B$  were  $n - 2, n - 1, n, n + 1,$  and  $n + 2$ . Figure 3 shows a sample set of slices belonging to



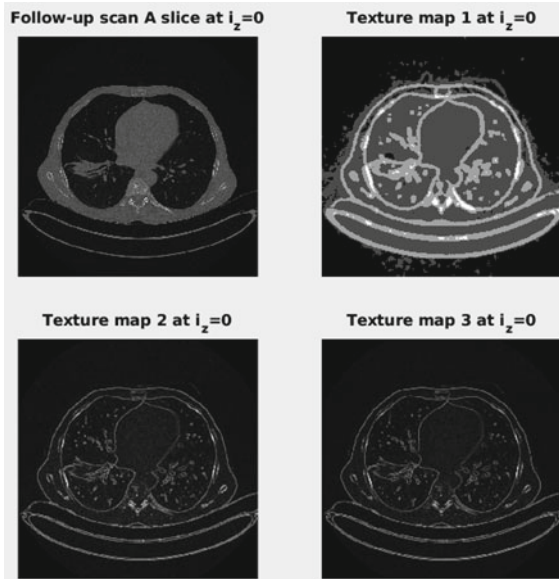
**Fig. 3** A sample set of follow-up scan  $B$  slices with corresponding texture maps when extended DBLM's search space size  $N_z = 5$

a search space belonging in follow-up scan  $B$ . Also, in Fig. 3, the slices belonging to the texture maps were shown. Similar to the follow-up scan  $B$ , the slices and texture maps of the follow-up scan  $A$  were shown in Fig. 4.  $I_{fA}^E$ ,  $I_{fB}^E$ ,  $I_{fA}^R$ ,  $I_{fB}^R$ ,  $I_{fA}^S$  and  $I_{fB}^S$  were the texture maps extracted from  $I_{fA}$  and  $I_{fB}$  3D arrays representing the entropy, range and standard deviation maps respectively.

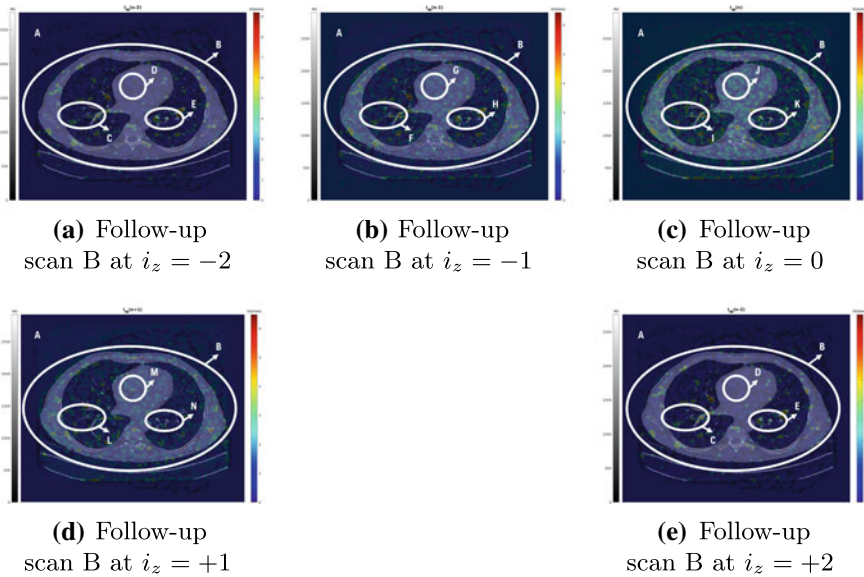
In order to illustrate the DVF estimation using the Extended DBLM, a sample set of slices belonging to the follow-up scan  $B$  are shown in Fig. 5. These slices are annotated with the labels **A** to **Q**, where, **A** belongs to the non-anatomical region and **B** belongs to the whole anatomical region in the slice locations  $i_z = -2$ ,  $i_z = -1$ ,  $i_z = 0$ ,  $i_z = +1$ , and  $i_z = +2$ . Similar to the sample set of the follow-up scan  $B$ , the sample slice of the follow-up scan  $A$  at the corresponding slice location  $i_z = 0$ , is shown in Fig. 6a with the regions annotated as **R**, **S**, and **T**. Figure 5 illustrates the estimated DVF overlaid on the origin of the displacement/motion activity in the slices of the follow-up scan  $B$ , whereas Fig. 6a shows the estimated DVF that describes the geometrical relationship of the corresponding slice in the follow-up scan  $A$ . The DVF overlaid on the slices in Figs. 5 and 6a used the magnitude of the motion vectors in its corresponding slice locations to highlight the origin of regularised displacements.

The displacement of the voxel blocks in the region **R**, i.e., the estimated motion vectors in Fig. 6a are heavily influenced by the regions **C**, **F**, **I**, **L**, and **O** highlighted in Fig. 6a. Similar to this, the DVF, i.e. motion vectors in the region **S** in Fig. 6a are heavily influenced by the regions **D**, **G**, **J**, **M**, and **P** in Fig. 5. Finally, the estimated DVF in the region **T** in Fig. 6a is heavily dependent on the regions **E**, **H**, **K**, **N**, and **Q** which are highlighted in Fig. 5. The DVF estimation using Extended DBLM was able to characterise the geometrical changes between the two follow-up scans as discussed in Sect. 2.2. However, in-terms of the clinical perspective, the results

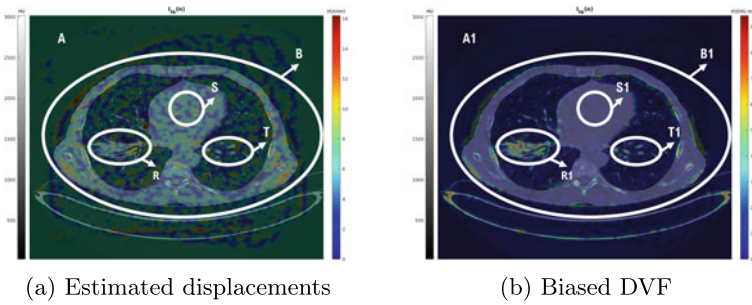




**Fig. 4** A sample follow-up scan A slice with corresponding texture maps



**Fig. 5** A sample set of follow-up scan B slices (a–e) with the regions of the activity overlaid when  $N_z = 5$



**Fig. 6** DVFs of extended DBLM in follow-up scan A slice at  $i_z = 0$

illustrated in Figs. 5 and 6a were insufficient to highlight areas of significance. Therefore, a visualisation technique was applied to the estimated DVF, i.e. motion vectors, to emphasise the regions of the activity. Since the DVF was biased with weights for enhanced visualisation unlike the original, where only the magnitude of the motion vectors was plotted, this was named as Biased DVF. Similar to the annotations in Fig. 6a, the regions in Fig. 6b were marked in the same locations except the labelling, where, to all those labels, a numeric '1' was used as suffix. To highlight the motion activity for the follow-up scan A, the absolute differences between the corresponding slice of the follow-up scan A and the search group slices of the follow-up scan B were calculated. Then, these absolute difference maps were then summed to become one combined map. The voxel intensities from the combined map were then multiplied to the magnitude of the motion vectors according to their corresponding locations, thus generating a Biased DVF for a slice in the follow-up scan A as shown in Fig. 6b. The region **S1** in Fig. 6b indicates there were no significant geometrical changes whereas the regions **R1** and **T1** indicates significant geometrical changes of higher magnitude as per the colour-bar beside it. For further verification, slice-wise SSIM and NMSE scores were calculated between the warped scan B with the ground-truth scan A. Table 1 provided the SSIM scores and the NMSE scores of the slices at every slice locations in the 3D data. From this Table, it was observed that the structural components of the estimated CT slice was very similar to the ground-truth CT slice given by the averaged SSIM score with the value of 0.996. In terms of the intensity based similarity, the averaged MSE and PSNR scores with values of  $\approx 0.003$  and  $\approx 46$  showed that the estimated CT slice was closer to the ground-truth. Also, from the other descriptive statistics such as range and standard deviation, it was observed that the DVF estimation was consistent with values closer to the mean and median of the performance metrics. Thus, the Extended DBLM was able to establish a geometrical relationship between two follow-up scans that had large coarser regions by using texture information as maps rather than GLCM features as a part of two-pass regularised disparity function without increasing the computational complexity. Along with that, the Extended DBLM provides a quantitative approach to measure the geometrical changes between two scans aiding the clinicians to characterise the tumour degradation or pulmonary fibrosis in-terms of volumetric and shape changes.

**Table 1** Performance evaluation of DVF estimation using extended DBLM

Slice-wise averaged descriptive statistics	SSIM	NMSE	PSNR
Mean	0.9960	0.0029	46.4434
Median	0.9959	0.0025	46.9835
Std.	0.0006	0.0008	1.0864
Range	0.0027	0.0024	3.2281

## 4 Conclusion

Quantified characterisation of the pulmonary fibrosis and other anomalies in lung CT have always been a challenge for clinicians. This paper has introduced a disparity-regularised block matching technique that uses both texture maps and voxel intensities as ISM to establish geometrical relationship between scans. Using two-pass distance and orientation based regularisation, this technique constrained the estimated motion vectors disallowing the unnatural and unrealistic deformations, particularly for the scan that has large coarser regions. With the image quality metrics, it was observed that the DVF was able to predict displacements that were closer to ground-truth. Thus, this technique aids in identifying and quantifying clinically significant geometrical changes in lung CT for SABR. Further investigations are yet to be conducted to deduce the suitable texture feature maps that could improve the robustness of this technique in identifying anomalies in the lung CT.

## References

1. Boersma, L.J., van den Brink, M., Bruce, A.M., Shouman, T., Gras, L., te Velde, A., Lebesque, J.V.: Estimation of the incidence of Late Bladder and Rectum Complications after High-Dose (70-78 Gy) Conformal Radiotherapy for Prostate Cancer, using Dose-Volume Histograms. *International Journal of Radiation Oncology\*Biophysics* **41**(1), 83–92 (1998). [https://doi.org/10.1016/S0360-3016\(98\)00037-6](https://doi.org/10.1016/S0360-3016(98)00037-6), <http://www.sciencedirect.com/science/article/pii/S0360301698000376>
2. Brock, K.: *Image Processing in Radiation Therapy. Imaging in Medical Diagnosis and Therapy*, CRC Press (2016), <https://books.google.co.uk/books?id=wVvRBQAAQBAJ>
3. Burnet, N.G., Thomas, S.J., Burton, K., Jefferies, S.J.: Defining the tumour and target volumes for radiotherapy. *Cancer Imaging* **4**, 153–161 (2004)
4. Cui, Z., Qi, W., Liu, Y.: A fast image template matching algorithm based on normalized cross correlation. *Journal of Physics: Conference Series* **1693**(1), 12163 (2020)
5. Dang, A., Kupelian, P.A., Cao, M., Agazaryan, N., Kishan, A.U.: Image-guided Radiotherapy for Prostate cancer. *Translational Andrology and Urology* **7**(3) (2018), <http://tau.amegroups.com/article/view/17960>
6. Gonzalez, R.C., Woods, R.E.: *Digital image processing*. Prentice Hall, Upper Saddle River, N.J. (2008), <http://www.amazon.com/Digital-Image-Processing-3rd-Edition/dp/013168728X>
7. Hayakawa, H., Shibata, T.: Block-matching-based motion field generation utilizing directional edge displacement. *Computers & Electrical Engineering* **36**(4), 617–625 (2010). <https://>

- [doi.org/10.1016/j.compeleceng.2008.11.017](https://doi.org/10.1016/j.compeleceng.2008.11.017), <https://www.sciencedirect.com/science/article/pii/S0045790608001146>, Signal Processing and Communication Systems
8. Huger, S., Graff, P., Harter, V., Marchesi, V., Royer, P., Diaz, J., Aouadi, S., Wolf, D., Peiffert, D., Noel, A.: Evaluation of the Block Matching deformable registration algorithm in the field of head-and-neck Adaptive Radiotherapy. *Physica Medica* **30**(3), 301–308 (2014). <https://doi.org/10.1016/j.ejmp.2013.09.001>, <http://www.sciencedirect.com/science/article/pii/S1120179713003967>
  9. Qian, X., Liu, G., Wang, H.: Texture based selective block matching algorithm for error concealment. In: 2007 IEEE International Conference on Multimedia and Expo. pp. 739–742 (2007). <https://doi.org/10.1109/ICME.2007.4284756>
  10. Rahmat, R., Harris-Birtill, D., Finn, D., Feng, Y., Montgomery, D., Nailon, W.H., McLaughlin, S.: Radiomics-led monitoring of non-small cell lung cancer patients during radiotherapy. In: Papież, B.W., Yaqub, M., Jiao, J., Namburete, A.I.L., Noble, J.A. (eds.) *Medical Image Understanding and Analysis*. pp. 532–546. Springer International Publishing, Cham (2021)
  11. Rigaud, B., Simon, A., Castelli, J., Lafond, C., Acosta, O., Haigron, P., Cazoulat, G., de Croisier, R.: Deformable image registration for radiation therapy: principle, methods, applications and evaluation. *Acta Oncologica* **58**(9), 1225–1237 (2019). <https://doi.org/10.1080/0284186X.2019.1620331>, pMID: 31155990
  12. Seferidis, V., Ghanbari, M.: Hierarchical motion estimation using texture analysis. In: 1992 International Conference on Image Processing and its Applications. pp. 61–64 (1992)
  13. Selvam, D.A.P., Laurenson, D.I., Nailon, W.H., McLaren, D.B.: Localised 3D disparity regularisation for improving contour propagation in Adaptive Radiotherapy. In: Işgum, I., Landman, B.A. (eds.) *Medical Imaging 2021: Image Processing*. vol. 11596, p. 115962U. International Society for Optics and Photonics, SPIE (2021). <https://doi.org/10.1117/12.2580574>
  14. Shepard, A.J., Wang, B., Foo, T.K., Bednarz, B.P.: A block matching based approach with multiple simultaneous templates for the real-time 2d ultrasound tracking of liver vessels. *Medical physics (Lancaster)* **44**(11), 5889–5900 (2017)
  15. Siciarz, P., Mccurdy, B., Alshafa, F., Greer, P., Hatton, J., Wright, P.: Evaluation of CT to CBCT non-linear dense anatomical block matching registration for prostate patients. *Biomedical Physics & Engineering Express* **4**(4), 045033 (Jun 2018). <https://doi.org/10.1088/2057-1976/aacada>
  16. Sipan, M., Susiki, N.S.M., Yuniarno, E.M.: Image block matching based on glcm (gray level co-occurrence matrix) texture feature on grayscale image auto coloring. In: 2017 International Seminar on Intelligent Technology and Its Applications (ISITIA). pp. 302–306 (2017). <https://doi.org/10.1109/ISITIA.2017.8124099>

# Identification of Melanoma Diseases from Multispectral Dermatological Images Using a Novel BSS Approach



Mustapha Zokay and Hicham Saylani

**Abstract** In this paper we propose a new approach to identify melanoma diseases by identifying the distribution of its main skin chromophores (melanin, oxyhemoglobin and deoxyhemoglobin) from multispectral dermatological images. Based on Blind Source Separation (BSS), our approach takes into account the shading present in most of the images. Assuming that the multispectral images have at least 4 spectral bands, it allows to estimate the distribution of each chromophore in addition to the shading without any a priori information, contrary to all existing methods that use 3 bands, i.e. RGB images. Indeed, the fact of neglecting the shading degrades their performance. To validate our method, we used a database of real multispectral dermatological images of skin affected by *melanoma* cancer. To measure our performance, in addition to the classical criterion of visually analyzing the estimated distributions with referring to the physiological knowledge of the disease, we proposed a new criterion that is based on our independence hypothesis. Using these two criteria, we could see that our approach is very efficient for the identification of melanoma.

**Keywords** Multispectral dermatological images · Chromophores · Melanin · Hemoglobin · Oxyhemoglobin · Deoxyhemoglobin · Shading · Blind source separation (BSS) · Melanoma

## 1 Introduction

The skin is the largest organ in the human body. It contains three main chromophores which are melanin, oxyhemoglobin and deoxyhemoglobin. The distribution of these chromophores is of great importance for dermatologists who use it for the identification and monitoring of skin diseases. An increasingly used technique that has

---

M. Zokay (✉) · H. Saylani

Laboratoire de Matériaux, Signaux et Systèmes et Modélisation Physique, Faculté des Sciences, Université Ibn Zohr, BP 8106 Cité Dakhla, Agadir, Morocco  
e-mail: [mustapha.zokay@edu.uiz.ac.ma](mailto:mustapha.zokay@edu.uiz.ac.ma)

H. Saylani

e-mail: [h.saylani@uiz.ac.ma](mailto:h.saylani@uiz.ac.ma)

proven to be effective for identifying the distribution of chromophores is multispectral imaging [1]. However, the direct use of multispectral images tends to give erroneous information on the distribution of chromophores. Indeed, the light intensity reflected by the skin does not only depend on these three chromophores but also on the geometry of the skin surface, called shading. The multispectral images obtained at different wavelengths can therefore be considered as mixtures of four constituents which are the three chromophores and the shading. Thus, if we consider these constituents as sources, this problem can be seen as a source separation problem, known as an inverse problem that belongs to the field of signal processing. The idea behind source separation is to estimate the sources exploiting only their mixture. Since it is performed without any a priori information, neither on the sources, nor on the mixing coefficients, it is called Blind Source Separation (BSS). It is easy to see that the BSS problem is an ill-posed inverse problem that admits an infinite number of solutions. Hence, it is essential to add hypotheses on the sources and/or on the mixing coefficients, which has led to 3 main families of BSS methods: *Independent Component Analysis* (ICA), *Sparse Component Analysis* (SCA) and *Non-negative Matrix Factorization* (NMF) (see [2] for more details). During the last decade, the use of BSS methods for non-invasive identification of chromophore distributions has attracted the interest of several researchers [3–9] who all adopted the same mixing model. Knowing that most of these researchers were interested in RGB images and thus in 3 wavelengths  $\lambda_i, i \in \{1, 2, 3\}$  which represent respectively the central wavelengths of the Blue, Green and Red bands, and if we note  $j \in \{1, 2, 3\}$  the index related to the chromophores and which represents respectively melanin, oxyhemoglobin and deoxyhemoglobin, then the classical mixing model is written:

$$I_{\lambda_i}(\mathbf{u}) = \sum_{j=1}^{j=3} a_{ij} \cdot S_j(\mathbf{u}) + p_d(\mathbf{u}) + n_i, \quad i \in \{1, 2, 3\}, \quad (1)$$

where

- $I_{\lambda_i}(\mathbf{u})$  is the logarithm inverse of the reflectance detected by the camera, at wavelength  $\lambda_i$ , at the pixel of coordinates  $(x, y) = \mathbf{u}$ ,
- $S_j(\mathbf{u})$  represents the chromophore of index  $j$ ,
- $a_{ij}$  represents the mixing coefficient which depends on the molar absorption coefficient of the chromophore  $S_j(\mathbf{u})$  and the light penetration depth in the skin at wavelength  $\lambda_i$ ,
- $p_d(\mathbf{u})$  represents the shading variation in the image,
- $n_i$  represents the characteristics of the sensor.

This mixing model has been adopted by all existing BSS methods [3–8], but they differ in the assumptions made about the chromophores and the procedure followed to reach the final objective, so that these methods can be grouped into two main classes. The first class includes the BSS methods that consider oxyhemoglobin and deoxyhemoglobin as a single source called hemoglobin [3–5, 9, 10]. Indeed, in [5, 10], the authors applied *Principal Component Analysis* (PCA) and then ICA on an

RGB image of the skin, assuming that the shading is constant throughout the image and that melanin and hemoglobin are independent. In [11], Madooei et al. proposed a new 2-D color chromaticity to eliminate shading using the geometric mean color, and then estimated the distributions of the two chromophores using *ICA*. In [9], Gong et al. proposed to estimate the distributions of both chromophores from an RGB image using *NMF*. In [3], Galeano et al. proposed to use a neural network-based system and then applied *NMF* to separate the melanin from the hemoglobin, but in their model they neglected the shading.

The second class includes the BSS methods which are based on a priori information on the absorption spectra of the three chromophores and the light penetration depth in the skin [6, 7]. The authors proposed to remove the shading and the specular reflection from the RGB image using respectively white paper and polarizers, then they relied on knowledge of absorption coefficients and estimates the light penetration depth in order to deduce an empirical mixing matrix to extract the distributions of the three chromophores. It should be noted, that the weakness of this family lies in the level of accurate estimation of light penetration depth into the skin.

In this paper, we propose a new method based on BSS to estimate the distribution of all the three main chromophores separately, in addition to the shading distribution which we consider as a full-fledged source, unlike all existing methods. Based on more realistic assumptions and applying to multispectral images with at least 4 spectral bands, our new method exploits the intrinsic properties of each chromophore. To validate our method we use a database of real multispectral dermatological images of skin affected by melanoma cancer disease [12]. For the performance measurement, in addition to the standard criterion that is based on the visual analysis of the three estimated chromophore distributions, we propose in this paper a new numerical criterion which is based on the measure of independence between the estimated distributions of melanin and hemoglobin. The rest of this paper is organized as follows. Section 2 presents our new method for estimating the distribution of the three chromophores and shading. Section 3 presents the results of the tests carried out followed by a last section devoted to a conclusion and perspectives for our work.

## 2 Proposed Method

The first idea behind our new method is to consider shading as a full-fledged source, in addition to the three sources of interest, which allows us to avoid the unrealistic assumption made by most existing methods that its contribution is the same at all wavelengths.<sup>1</sup> However, as the number of sources involved becomes equal to 4 we are interested in this paper for the case where we have multispectral images with at

---

<sup>1</sup> Indeed, we found from the experimental curve obtained by *PCA* used in [13] that the contribution of shading ( $p_d$ ) is not equal to 1 in all mixtures.

least 4 bands, said case determined.<sup>2</sup> From Eq. (1), we can see that the term  $n_i$  does not give any information on chromophore distributions. Thus, as in [5], we begin by eliminating it from our mixtures based on the following hypothesis (**H1**).

(**H1**): There is at least one pixel in the 4 images where the concentrations of the three chromophores and shading are all zero, i.e.:

$$\exists \mathbf{u}/n_i = \min(I_{\lambda_i}(\mathbf{u})). \quad (2)$$

So, the new mixture model is written:

$$X_i(\mathbf{u}) = \sum_{j=1}^{j=3} a_{ij} \cdot S_j(\mathbf{u}) + a_{i4} \cdot S_4(\mathbf{u}), \quad i \in [1, 4] \quad (3)$$

where  $X_i(\mathbf{u}) = I_{\lambda_i}(\mathbf{u}) - n_i$ ,  $S_4(\mathbf{u}) = p_d(\mathbf{u})$  and  $a_{i4} \in \mathbb{R}$ .

In the same way as with all BSS methods, we produce a new set of 1D mixtures (vectors), which we note  $X_i(v)$ , from the 2D mixtures (images)  $X_i(\mathbf{u})$  by concatenating the rows of the latter. We then have:

$$X_i(v) = \text{vec}(X_i(\mathbf{u})) = \sum_{j=1}^{j=4} a_{ij} \cdot S_j(v), \quad i \in [1, 4] \quad (4)$$

The second idea behind our method is to treat mixtures in two steps, unlike all existing methods that treat all mixtures at the same time. Indeed, we start by treating only two mixtures that contain only melanin and shading in order to separate them first, and then we eliminate their contribution from the other two mixtures remaining to keep only oxyhemoglobin and deoxyhemoglobin. These last two chromophores are then separated in a last step. These three steps of our method are detailed below.

### Step 1: Separation of sources $S_1(v)$ and $S_4(v)$

In this step we exploit the properties of each chromophore concerning spectral absorption as a function of the wavelength. Indeed, based on data published in [14], we found that light absorption at wavelengths greater than 620 nm is dominated by melanin, so the absorption coefficients  $a_{ij}$  of oxyhemoglobin and deoxyhemoglobin are all negligible, i.e. we have  $a_{32} = a_{33} = 0$  and  $a_{42} = a_{43} = 0$ . Thus, the mixtures corresponding to the red and infrared bands can be re-written as follows:

$$\begin{cases} X_3(v) = a_{31} \cdot S_1(v) + a_{34} \cdot S_4(v) \\ X_4(v) = a_{41} \cdot S_1(v) + a_{44} \cdot S_4(v) \end{cases} \quad (5)$$

---

<sup>2</sup> This is the case where we have as many mixtures as sources. On the other hand, in the case where we have more mixtures than sources, called the over-determined case, we can easily return to the determined case by applying a *PCA*.



We can re-write the equation system (5) in a matrix form as follows:

$$\mathbf{X}(v) = \mathbf{A} \cdot \mathbf{S}(v), \quad (6)$$

where  $\mathbf{X}(v) = [X_3(v), X_4(v)]^T$ ,  $\mathbf{S}(v) = [S_1(v), S_4(v)]^T$  and  $\mathbf{A} = \begin{pmatrix} a_{31} & a_{34} \\ a_{41} & a_{44} \end{pmatrix}$ .

In this first step we assumed that the two sources  $S_1(v)$  and  $S_4(v)$  are independent, in which case we can use one of the *ICA* methods to separate them. We have opted here for the *AMUSE* method [15] for its simplicity since it exploits only the second order statistics of the signals. Indeed, the working hypotheses of this method are the following.

**(H2):** The sources  $S_j(v)$  are *auto-correlated* and *mutually uncorrelated*, i.e.:

$$\forall \tau, \begin{cases} E[S_j(v) \cdot S_j(v - \tau)] \neq 0, & j \in \{1, 4\} \\ E[S_1(v) \cdot S_4(v - \tau)] = E[S_1(v)] \cdot E[S_4(v - \tau)] \end{cases} \quad (7)$$

**(H3):** The *condition of identifiability* for the method is verified, i.e.:

$$\exists \tau \neq 0 / \frac{E[S_1(v) \cdot S_1(v - \tau)]}{E[S_1^2(v)]} \neq \frac{E[S_4(v) \cdot S_4(v - \tau)]}{E[S_4^2(v)]}. \quad (8)$$

The method *AMUSE* allows us to estimate the separation matrix  $\mathbf{A}^{-1}$  to a permutation matrix  $\mathbf{P}$  and a diagonal matrix  $\mathbf{D}$  [15]. By noting this matrix  $\mathbf{C} = \mathbf{PDA}^{-1}$ , we obtain finally the source matrix  $\mathbf{S}(v)$  with the same indeterminations as follows:

$$\mathbf{C} \cdot \mathbf{X}(v) = (\mathbf{PDA}^{-1}) \cdot (\mathbf{AS}(v)) = \mathbf{PDS}(v). \quad (9)$$

Indeed, by noting  $\mathbf{PDS}(v) = \mathbf{Y}(v) = [Y_1(v), Y_4(v)]$  and omitting the permutation<sup>3</sup> we have:

$$Y_j(v) = \alpha_j \cdot S_j(v), \quad j = 1, 4, \quad (10)$$

where the  $\alpha_j$  are the elements constituting the diagonal of the matrix  $\mathbf{D}$ .

### Step 2: Removal of $S_1(v)$ and $S_4(v)$ sources from mixtures

The goal of this step is to eliminate the contributions of the sources estimated  $S_1(v)$  and  $S_4(v)$  from the mixtures  $X_1(v)$  and  $X_2(v)$ . For this we exploit the following independence hypothesis.

**(H4):**  $S_i(v)$  and  $S_j(v)$  are *mutually uncorrelated instantaneously*, for  $i \in \{2, 3\}$  and  $j \in \{1, 4\}$ , i.e.:

$$E[\tilde{S}_i(v) \cdot \tilde{S}_j(v)] = E[\tilde{S}_i(v)] \cdot E[\tilde{S}_j(v)] = 0, \quad \forall (i, j) \in \{2, 3\} \times \{1, 4\} \quad (11)$$

<sup>3</sup> Indeed, the permutation matrix  $\mathbf{P}$  can be identified based on the visual analysis since the shading source  $S_4(v)$  is easily differentiable.

where  $\tilde{S}_i(v)$  are the centered versions<sup>4</sup> of the sources  $S_i(v)$ . By denoting respectively  $\tilde{X}_i(v)$  and  $\tilde{Y}_j(v)$  the centered versions of the signals  $X_i(v)$  and  $Y_j(v)$ , and by exploiting the relations (10) and (11) we can write:

$$E[\tilde{X}_i(v) \cdot \tilde{Y}_j(v)] = E \left[ \left( \sum_{k=1}^{k=4} a_{ik} \cdot \tilde{S}_k(v) \right) \cdot (\alpha_j \cdot \tilde{S}_j(v)) \right] \quad (12)$$

$$= a_{ij} \cdot \alpha_j \cdot E[\tilde{S}_j^2(v)] \quad (13)$$

On the other hand, according to the relation (10) we have:

$$E[\tilde{Y}_j^2(v)] = \alpha_j^2 \cdot E[\tilde{S}_j^2(v)], \quad j = 1, 4. \quad (14)$$

Thus, by exploiting the relations (13) and (14) we can generate two new mixtures  $Z_i(v)$  ( $i = 1, 2$ ) which contain only the sources  $S_2(v)$  and  $S_3(v)$  as follows:

$$Z_i(v) = X_i(v) - \frac{E[\tilde{X}_i(v) \cdot \tilde{Y}_1(v)]}{E[\tilde{Y}_1^2(v)]} \cdot Y_1(v) - \frac{E[\tilde{X}_i(v) \cdot \tilde{Y}_4(v)]}{E[\tilde{Y}_4^2(v)]} \cdot Y_4(v) \quad (15)$$

$$= X_i(v) - \frac{a_{i1}}{\alpha_1} \cdot (\alpha_1 \cdot S_1(v)) - \frac{a_{i4}}{\alpha_4} \cdot (\alpha_4 \cdot S_4(v)) \quad (16)$$

$$= a_{i2} \cdot S_2(v) + a_{i3} \cdot S_3(v) \quad (17)$$

### Step 3: Separation of Sources $S_2(v)$ and $S_3(v)$

The goal of this step is to separate the remaining  $S_2(v)$  and  $S_3(v)$  sources that represent oxyhemoglobin and deoxyhemoglobin respectively, and this time by treating the new mixtures  $Z_1(v)$  and  $Z_2(v)$  provided by the previous step. As these two sources are dependent, which is why most of the existing methods fail to separate them, we have opted in this paper for a new solution based on the exploitation of their sparsity.<sup>5</sup> Here is our working hypothesis for this step.

**(H5):** For each source, there is at least one spatial area over which it is active while the other source is inactive, i.e.:

$$\begin{cases} \exists V_1 / \forall v \in V_1, S_2(v) = 0 \text{ and } S_3(v) \neq 0 \\ \exists V_2 / \forall v \in V_2, S_3(v) = 0 \text{ and } S_2(v) \neq 0 \end{cases} \quad (18)$$

There are several BSS methods that exploit this sparsity assumption to achieve separation. We opted here for the *TEMPROM* method proposed in [16] for its simplicity. This method consists of identifying the single-source areas  $V_1$  and  $V_2$  at first, and then calculating the ratio between the mixtures  $Z_2(v)$  and  $Z_1(v)$  on these areas at a second time, which ultimately allows to estimate the matrix of mixture involved.

<sup>4</sup> i.e.:  $\tilde{S}_i(v) = S_i(v) - E[S_i(v)]$ .

<sup>5</sup> A source is said to be sparse in a given representation domain if some of its samples are zero in this domain.

Indeed, by exploiting the two equations of the system (18) and the relation (17) we obtain:

$$\begin{cases} \forall v \in V_1, \frac{Z_2(v)}{Z_1(v)} = \frac{a_{23} \cdot S_3(v)}{a_{13} \cdot S_3(v)} = \frac{a_{23}}{a_{13}} = r_1 \\ \forall v \in V_2, \frac{Z_2(v)}{Z_1(v)} = \frac{a_{22} \cdot S_2(v)}{a_{12} \cdot S_2(v)} = \frac{a_{22}}{a_{12}} = r_2 \end{cases} \quad (19)$$

Finally, by exploiting the relations (17) and (19) we obtain:

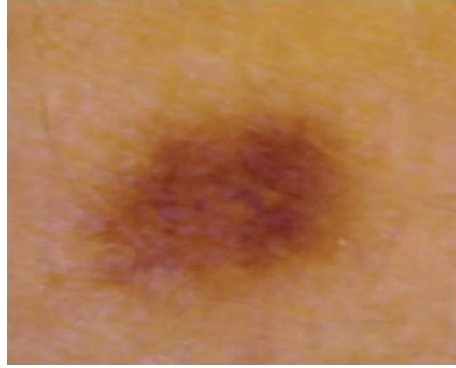
$$\begin{cases} r_1 \cdot Z_1(v) - Z_2(v) = (r_1 \cdot a_{12} - a_{22}) \cdot S_2(v) = \alpha_2 \cdot S_2(v) = Y_2(v) \\ r_2 \cdot Z_1(v) - Z_2(v) = (r_2 \cdot a_{13} - a_{23}) \cdot S_3(v) = \alpha_3 \cdot S_3(v) = Y_3(v) \end{cases} \quad (20)$$

where  $\alpha_2$  and  $\alpha_3$  are scalars.

### 3 Results

In this section, we measure the performances of our method using a database of *melanoma skin cancer* patients which is an open access database [12]. It is recalled that our method makes it possible to estimate the distributions of oxyhemoglobin and deoxyhemoglobin separately in addition to those of melanin and shading contrary to all existing methods (as mentioned in the introduction). So, in absolute terms we cannot compare our performance with any of these methods. However, since most of these methods estimate melanin and hemoglobin (which is a mixture of oxyhemoglobin and deoxyhemoglobin), we can limit ourselves in the comparison to these two chromophores. For this, we opted for a comparison with the methods proposed in [9, 11] because they are accessible for testing, unlike most of the existing methods [4, 6, 7]. As for the performance measurement criteria, in addition to the classical visual criterion which is a subjective criterion, we propose in this paper a new numerical criterion. In fact, for dermatologists, melanoma is characterized by a high distribution of melanin, an average distribution of deoxyhemoglobin and a very low distribution of oxyhemoglobin compared to healthy skin [17]. To support this subjective criterion, which is the most used in the literature [4, 6, 7, 9], we use a second numerical criterion, which allows us to check to what extent the independence hypothesis we have assumed is satisfied by the estimated chromophores, since this hypothesis is the most used by researchers [5, 10, 11]. We then define our new performance measurement criterion, which we note as  $C_{Ind}$ , by exploiting the assumption of statistical independence at order 4 between melanin and hemoglobin, as follows:

$$C_{Ind} = \frac{1}{2} (C_{13} + C_{31}), \quad (21)$$



**Fig. 1** Treated multispectral dermatological image

where

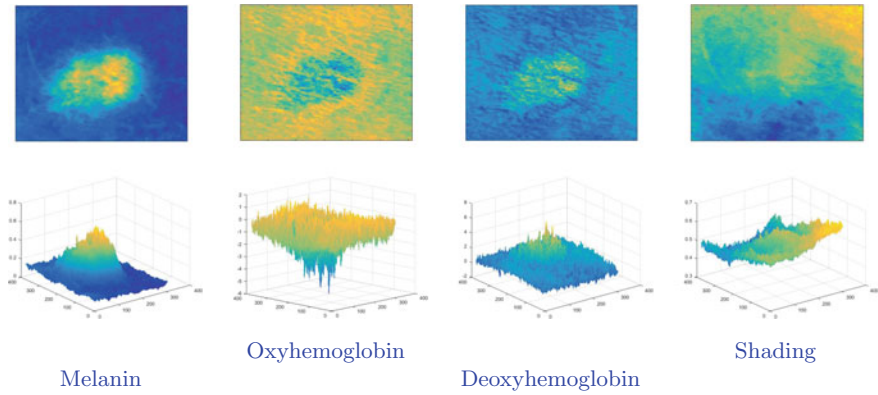
$$C_{13} = 20 \log_{10} \left( \left| E[\tilde{Y}_1(u)\tilde{Z}_1(u)^3] \right|^{-1} \right) \text{ and } C_{31} = 20 \log_{10} \left( \left| E[\tilde{Y}_1(u)^3\tilde{Z}_1(u)] \right|^{-1} \right) \quad (22)$$

We recall that the estimates  $\tilde{Y}_1(u)$  and  $\tilde{Z}_1(u)$  are respectively the centered versions of the melanin and hemoglobin distributions.

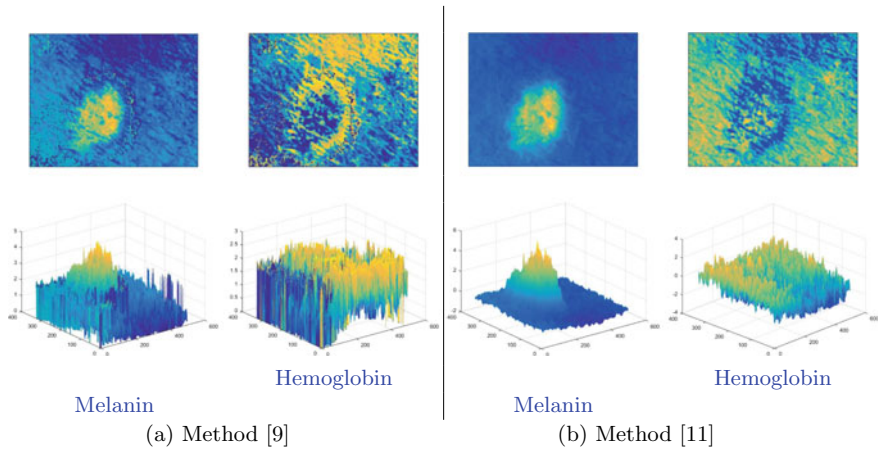
The database contains 30 multispectral images [12]. Knowing that we have tested our method on some images and that the results obtained are similar, we limit ourselves here to present the results of a single melanoma image, due to lack of space. This image is shown in Fig. 1.

For the methods [9, 11], we used respectively the algorithm FastICA [18] and MU [19]. Two-dimensional (2D) and three-dimensional (3D) representations of the estimated distributions of each of the three chromophores in addition to the shading using our method are given in Fig. 2. These 2D and 3D representations are grouped by column for each chromophore. The 2D and 3D representations of the estimated melanin and hemoglobin distributions using the methods [9, 11] are grouped in Fig. 3.

Figure 2 shows the good performance of our method for the estimation of the distribution of each of the three chromophores in addition to the shading. Indeed, we see that in the melanoma area we have a high distribution of melanin and a relatively high distribution of deoxyhemoglobin compared to oxyhemoglobin as shown in the 3D representations of these chromophores. This is fully consistent with the physiological knowledge that characterizes melanoma, since *naevus* disease, which generally has a melanin distribution similar to that of *melanoma*, is instead characterized by an oxyhemoglobin distribution that is close to that of deoxyhemoglobin [17]. On the other hand, from Fig. 3, we see that the melanin distribution estimated by the methods [9, 11] is similar to that estimated by our method, which means that these two methods [9, 11] also suspects melanoma. Nevertheless, its estimated hemoglobin distribution (which can be seen as a mixture of oxyhemoglobin, deoxy-



**Fig. 2** Estimated distributions (2D and 3D) of the three chromophores and shading using *our method*



**Fig. 3** Estimated distributions (2D and 3D) of melanin and hemoglobin using: **a** method [9], **b** method [11]

hemoglobin, and shading) does not allow to decide on the nature of the disease, since the distinction between melanoma and nevus can be made only by estimating the distributions of oxyhemoglobin and deoxyhemoglobin separately. All these findings, which were made on the basis of physiological knowledge of the melanoma disease, are in perfect agreement with the results obtained using our new numerical criterion, denoted  $C_{Ind}$  and defined by the relation (21), which are presented in Table 1. In this table we provide the mean and standard deviation of  $C_{Ind}$ , denoted  $\bar{C}_{Ind}$  and  $\sigma$ , obtained on 10 different images from the database [12].

**Table 1** Mean and standard deviation of  $C_{Ind}$  in dB

	Method [9]	Method [11]	Our method
$\overline{C}_{Ind}$ (dB)	-2.05	27.45	<b>38.44</b>
$\sigma$ (dB)	3.66	6.20	<b>4.33</b>

From Table 1, we see that our method is much better than the other methods since the value obtained for  $\overline{C}_{Ind}$  using our method is larger than that obtained using the methods [9, 11]. We also find that the [9] method has poor performance compared to our method and the method [11] and this can be explained by the infinite solution problem posed by NMF. Since this criterion is based on a measure of independence, we deduce that our method provides estimates of melanin and hemoglobin distributions in output that are much more independent than those provided by the methods [9, 11].

## 4 Conclusion

In this paper we have proposed a new approach which aims to identify the melanoma diseases by identifying the distribution of its main skin chromophores (melanin, oxyhemoglobin and deoxyhemoglobin) from multispectral dermatological images. The key idea of our approach is to take into account the shading, considering it as a full-fledged source, and the three chromophores, which leads to an instantaneous linear mixture model with four sources rather than two or three sources, as is the case for all existing methods. The results of all the tests carried out, using a database of real multispectral images of skin affected by a skin cancer of the type melanoma, have shown that our approach is very efficient using the classical criterion based on visual analysis than our new independence criterion. In terms of perspective, it would be interesting to test our method on other multispectral dermatological image databases of other skin diseases.

## References

1. Jacques, S.L., Samatham, R., Choudhury, N.: Rapid spectral analysis for spectral imaging. *Biomedical optics express* 1(1), 157–164 (2010)
2. Comon, P., Jutten, C.: *Handbook of Blind Source Separation, Independent Component Analysis and Applications* (02 2010)
3. Jolivot, R., Marzani, F., et al.: Quantification of melanin and hemoglobin in human skin from multispectral image acquisition: use of a neuronal network combined to a non-negative matrix factorization. *Applied and Computational Mathematics, special issue on Applied Artificial Intelligence and Soft Computing* 11(2), 257–270 (2012)

4. Liu, Z., Zerubia, J.: Melanin and hemoglobin identification for skin disease analysis. In: 2013 2nd IAPR Asian Conference on Pattern Recognition. pp. 145–149. IEEE (2013)
5. Ojima, N., Akazaki, S., Hori, K., Tsumura, N., Miyake, Y.: Application of image-based skin chromophore analysis to cosmetics. *Journal of Imaging Science and Technology* 48(3), 222–226 (2004)
6. Spigulis, J., Oshina, I.: Snapshot rgb mapping of skin melanin and hemoglobin. *Journal of biomedical optics* 20(5), 050503 (2015)
7. Spigulis, J., Oshina, I., Berzina, A., Bykov, A.: Smartphone snapshot mapping of skin chromophores under triple-wavelength laser illumination. *Journal of Biomedical Optics* 22(9), 091508 (2017)
8. Kuzmina, I., Diebele, I., Asare, L., Kempele, A., Abelite, A., Jakovels, D., Spigulis, J.: Multispectral imaging of pigmented and vascular cutaneous malformations: the influence of laser treatment. In: *Laser Applications in Life Sciences*. vol. 7376, p. 73760J. International Society for Optics and Photonics (2010)
9. Gong, H., Desvignes, M.: Hemoglobin and melanin quantification on skin images. In: *International Conference Image Analysis and Recognition*. pp. 198–205. Springer (2012)
10. Mitra, J., Jolivot, R., Vabres, P., Marzani, F.S.: Source separation on hyperspectral cube applied to dermatology. In: *Medical Imaging 2010: Computer-Aided Diagnosis*. vol. 7624, p. 76243I. International Society for Optics and Photonics (2010)
11. Madooei, A., Drew, M.: A Bioinspired Color Representation for Dermoscopy Image Analysis, pp. 23–66 (09 2015)
12. Lézoray, O., Revenu, M., Desvignes, M.: Graph-based skin lesion segmentation of multispectral dermoscopic images. In: *International Conference on Image Processing (IEEE)*. pp. 897–901 (2014)
13. Eguizabal, A., Laughney, A.M., García-Allende, P.B., Krishnaswamy, V., Wells, W.A., Paulsen, K.D., Pogue, B.W., Lopez-Higuera, J.M., Conde, O.M.: Direct identification of breast cancer pathologies using blind separation of label-free localized reflectance measurements. *Biomedical optics express* 4(7), 1104–1118 (2013)
14. Van Gemert, M., Jacques, S.L., Sterenborg, H., Star, W.: Skin optics. *IEEE Transactions on biomedical engineering* 36(12), 1146–1154 (1989)
15. Tong, L., Liu, R.W., Soon, V.C., Huang, Y.F.: Indeterminacy and identifiability of blind identification. *IEEE Transactions on circuits and systems* 38(5), 499–509 (1991)
16. Abrard, F., Deville, Y.: A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. *Signal processing* 85(7), 1389–1403 (2005)
17. Lihacova, I.: Evaluation of skin oncologic pathologies by multispectral imaging methods. Ph.D. thesis (07 2015)
18. Hyvärinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. *Neural computation* 9(7), 1483–1492 (1997)
19. Lee, D., Seung, H.S.: Algorithms for non-negative matrix factorization. *Advances in neural information processing systems* 13 (2000)

# 2.5D Lightweight Network Integrating Multi-scale Semantic Features for Liver Tumor Segmentation



Yilin You , Zhengyao Bai , Yihan Zhang , and Jiajin Du 

**Abstract** One critical research area in the development of a computer-aided diagnosis system for liver cancer is efficient and automatic segmentation of lesion from CT scans. To overcome this issue, we investigated a 2.5D lightweight liver tumor segmentation by fusing the multi-scale semantic features, named MAA-Net. Our framework enhanced the information interaction between the input 2.5D stacked slice via introducing parallel convolution and increasing the knowledge weight of the lesion channel in different receptive fields. To ease the shortage of missed detection of tumors, MAA-Net fused the hierarchical semantic information extracted from the encoder. Moreover, we evaluated our MAA-Net on LiTS2017 and 3DIRCADb datasets. Extensive experiments shows the proposed method outperforms the others on both accuracy and total number of calculation. Specifically, our approach can improve liver tumor segmentation tasks by 2.4%, while reducing amount of parameters by 57.5%. Both quantitative and qualitative results illustrated the MAA-Net can effectively address with the limitation of small tumors, and some tumors are on the edge.

**Keywords** Liver tumor segmentation · Multi-scale feature fusion · Attention mechanism · U-Net

## 1 Introduction

Primary liver cancer seriously endangers the health of patients. Experienced doctors diagnose and treat patients by focusing on the location, shape, and size of lesions in CT scans. However, the contrast between lesion and adjacent soft tissues is low, while the shape of liver tumors is highly variable. Thus, it's tough to define the lesion boundary, which makes it impossible to achieve accurate tumor segmentation.

In response to the above issues, many sophisticated approaches have reached varying degrees of success. As an end to end Fully Convolutional Network (FCN)

---

Y. You · Z. Bai (✉) · Y. Zhang · J. Du  
School of Information Science and Engineering, Yunnan University, Kunming 650500, China  
e-mail: [baizhy@ynu.edu.cn](mailto:baizhy@ynu.edu.cn)



[1] occurred, the deep learning to pixel-level segmentation of lesion has become mainstream. Meanwhile, U-Net [2] used the symmetric structure and unique skip connections for feature integration. FSF U-Net [3] adopted residual fusion and global feature selection unit to predict the location of liver tumor. AHC-Net [4] considered the characteristics of volumetric CT scans, designed a cascade network and combined soft and hard attention to get exact target bounding boxes. H Dense U-Net [5] adopted a pre-trained model to get rough liver segmentation, and through 3D Dense U-Net to detect features between slices. Compared with the 2D network, 3D network can better capture the detailed spacial knowledge and achieve faster convergence speed. However, because of the complexity of 3D architecture and heavy computational burden, Triplanar FCN [6] mixed three 2D neural networks to segment the liver from the lateral, coronal, and sagittal planes, which effectively used multi-dimensional features. Hy\_CompNet [7] used 2D CompNet to segment the liver and large tumors to balance computing resources and accuracy. Considering the high similarity of blood vessels and lesions, 3D CompNet for small missed tumors segmentation.

Combining the characteristics of liver CT scans and U-Net architecture, the current obstacles for lesion segmentation are: (1). The gray value between tumors and adjacent tissues is resemble, while the variability of the location and shape leads to lower tumor segmentation results than liver; (2). The 2D network can't fully adopt the spacial information between slices within the sample, and 3D structure would consume too much computing resources.

Addressing with the above issues is what we exact concentrated on this paper. We proposed a 2.5D lightweight global semantic feature integration and multi-scale encoder-decoder liver tumor segmentation network, called MAA-Net (Multi-scale Attention Aware-Net). Overall, the main contributions of this work includes: (1). Extract features through the InceptionV3 with enhanced channel attention; (2). Employ a dual-feature fusion block in the skip connections to focus on the tiny features and suppress irrelevant information; (3). Acquire the inter-slice spatial information via 2.5D network structure.

## 2 Related Work

### 2.1 Inception Architecture

InceptionV3 [8] utilized asymmetric convolution in series to decompose the original convolution kernel. While ensuring the accuracy, the InceptionV3 saved 33% of the calculation of InceptionV2. Currently, many academicians employed the Inception to segment medical image. Reference [9] considered using max pooling and up-sampling in the lung nodules segmentation would affect the resolution of the feature map. Thus, they adopted a promoted multi-branch up and down-sampling to perceive local features of various ranges. RIU-Net [10] replaced the original convolution operation in the U-Net with InceptionV3 for medical image segmentation.

## 2.2 Residual Network

Residual structures [11] are effective in avoiding gradient vanishing and exploding with the increasing of network depth. Meanwhile, residual connections can fortify the combined information between the upper and lower layers. Squeeze and Excitation Network [12] utilized global average pooling and two fully connected layers to perform re-calibration of dynamic features. Ameliorating the representation ability of the network. Convolutional block attention module [13] adopted a serial architecture. It first used the channel attention vector to rectify the shallow feature map, then obtained final blended feature map through spatial attention. To ease the burden of computations, [14] compared the network parameters of SE\_Net, designed three variant experiments and added a band matrix to achieve 1D convolution. Ultimately, the Efficient Channel Attention block (ECA) through a function to implement adaptive cross-channel with the least amount of parameters.

## 2.3 2.5D Network Architecture

Directly using volumetric CT scans as input will bring redundant calculation, while slice image may waste of too much contextual information. Therefore, it is a good compromise to consider adopting a 2.5D network structure. Li and Bai [15] designed a 2.5D convolutional network for multi-site lesion detection. A CT slice of multi-lesion type is a continuous tomographic image, and only marks its key slice in the database. We can see from the continuity that the adjacent slices of a slice are spacial related. So the input of the network is with a CT slice group formed by stacking the annotated central slice and its adjacent slice carrying contextual information.

# 3 Method

## 3.1 MAA\_Net

As shown in Fig. 1, multi-scale features were first extracted by the asymmetric convolutional layers, named global context aware (GCA\_E). We embed ECA unit in each GCA module intends to notice the inter-slice information between distinct channels. Among them, we employed four dual-feature fusion blocks (DFB) through skip connection. Then we used the decoder to locate regions of interest, and added deep supervision to enhance the network transmission. Since the varying size of lesions, using a settled convolution kernel ineluctably leads to a fixed receptive field. Thus, MAA-Net designed a hierarchical structure to extract contextual information.

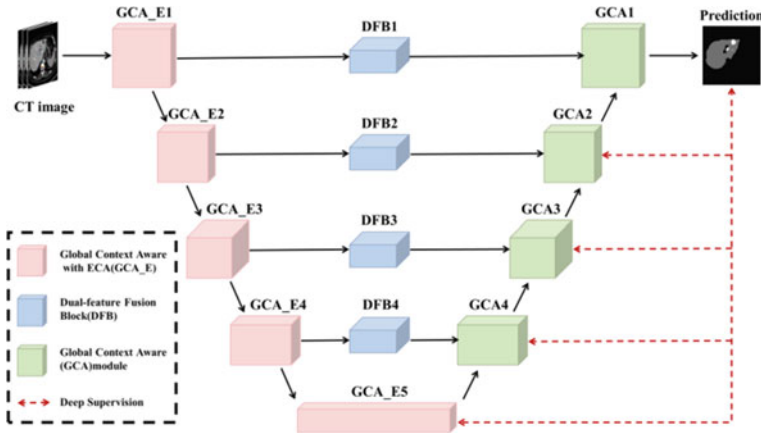


Fig. 1 Network structure

### 3.2 Encoder-Decoder Structure

Inspired by nCovSegNet [16], the encoder of MAA-Net employed InceptionV3 as the backbone for feature extraction. Figure 2 demonstrates the structure of the GCA module. To get a large receptive field without increasing the amount of parameters, the GCA has five parallel branches which consists different convolution layers.

In the expansive path, we adopted a parallel decoder with residual. As shown in Fig. 2, by aggregating semantic information between multi-level. The encoder can reduce computational redundancy while processing complex and diverse features.

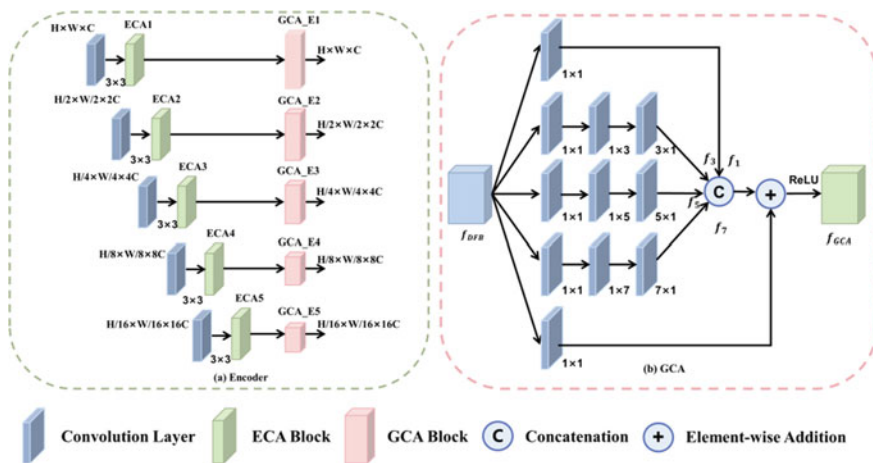


Fig. 2 Multi-branch parallel encoder and decoder

Moreover, 2D input images can't fully use the inter-slices information, 3D scans, on the contrary, may waste of time and computation. Therefore, taking 2.5D form as the input of the model by stacking adjacent slices to capture the spacial knowledge. Wardhana [17] compared the Dice and Hausdorff distance of stacked consecutive 3, 5, 7 and 9 slices, then concluded that three slices can provide the most association information without distorting tumors.

### 3.3 Dual-Feature Fusion Module

To integrate the high-level semantic information and low level detailed edge operation is beneficial for early screening and missed detection. Coordinate Attention (CA) [18] through the decomposition channel attention to 2 one-dimensional encoded features and aggregating in distinct directions to achieve multi-scale semantic feature fusion. The CA mechanism first obtains long distance dependencies along one direction while retaining precise location information with the other spatial direction. Thus, we added DFB modules in the skip connection of each layer to alleviate the loss of deep details and recovering spatial information. Specifically, as shown in Fig. 3, we considered both high-to-low and low-to-high data flows to gain the cross channel relationship and location information. Following, we divided the global average pooling into 2 steps, by using different pooling kernel to get feature maps.

We used the previously feature maps to concat along the spacial dimension, and transformed the bottleneck layer then activated to achieve information interaction. After fusing the hierarchical features, we separated the shared layers along the spatial dimension and utilized the sigmoid function to obtain the sum of attention vectors. Ultimately, we operated on the vector and input with Hadamard product. Mathematically, the description can be expressed as:

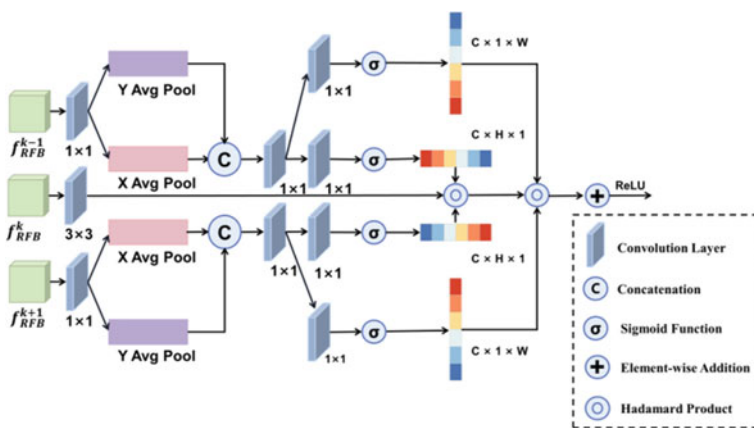


Fig. 3 Multi-scale semantic feature fusion module

$$f = \delta \left( F_1 \left( \left[ \frac{1}{W} \sum_{0 \leq i \leq W} x_c(h, i), \frac{1}{H} \sum_{0 \leq j \leq H} x_c(j, w) \right] \right) \right) \quad (1)$$

$$y_c(i, j) = x_c(i, j) \times \sigma(F_h(f_h)) \times \sigma(F_w(f_w)) \quad (2)$$

where  $z_c^h \in \mathbb{R}^{C \times H \times 1}$ ,  $z_c^w \in \mathbb{R}^{C \times 1 \times W}$ ,  $z_c$  are the embedded feature maps in the different direction. Let  $x_c \in \mathbb{R}^{C \times H \times W}$ , defined as the input of channel  $C$ , where  $C, H, W$  are the number of channels, height and width of the feature map, respectively. And  $[\cdot]$  demotes the splicing operation,  $F_1, F_h, F_w$  represent the bottleneck layer,  $\delta, \sigma$  are activation function.

### 3.4 Loss Function

The cross-entropy loss function frequently employed to measure the similarity between the prediction and the ground truth in semantic segmentation. The smaller loss indicates stronger robustness of the model. Its mathematical definition is:

$$L_{\text{cross}} = \frac{1}{N} \sum_i L_i = -\frac{1}{N} \sum_i \sum_{c=1}^M y_{ic} \log p_{ic} \quad (3)$$

where,  $M$  means the number of classes,  $y_{ic}$  is a sign function, which takes one when the classification is correct;  $p_{ic}$  represents the probability that the sample  $i$  belongs to  $c$ . Since lesions account for a relatively small proportion of entire CT scans, adopting a single loss function may cause in a severe class disequilibrium. By adding the Dice loss function to balance the proportion of tumors, and is defined as

$$L_{\text{Dice}} = 1 - \frac{2 \sum_j^N p(i, j) q(i, j)}{\sum_j^N p^2(i, j) + \sum_i^N q^2(i, j)} \quad (4)$$

where,  $N$  is the sum of voxels in the whole slice,  $i$  and  $j$  denotes the classes  $i$  and  $j$ ,  $p(i, j)$  and  $q(i, j)$  is defined as the probability that the voxels  $j$  belongs to the category  $i$  in the prediction and label, respectively. Therefore, the overall loss is

$$L = L_{\text{Dice}} + \gamma L_{\text{cross}} \quad (5)$$

where the weight  $\gamma$  set to 0.5 in the experiment.

**Table 1** The relevant information of LiTS 2017 and 3DIRCADb

Database	Training	Testing	Size	Flat resolution (mm)	Slice thickness (mm)	Depth
LiTS17	131	70	$512 \times 512$	0.55–1.0	0.45–6.0	42–1016
3DIRCADb	–	20	$512 \times 512$	0.57–0.87	1.6–4.0	74–260

## 4 Experiments

### 4.1 Datasets

To evaluate the proposed method, we employed two publicly available datasets, LiTS2017 [19] and 3DIRCADb [20]. Meanwhile, the relevant information is shown in Table 1. The experiment randomly selected 70 and 30% of the LiTS2017 samples to construct training and validation sets, and tested on the 3DIRCADb. We unified the Hounsfield unit to  $[-200, 200]$  and normalized the pixel value to reduce the influence of irrelevant background.

### 4.2 Setting Details

We implemented our method based on Pytorch and completed training on NVIDIA GeForce RTX2080Ti GPU. We used Adam as the optimization algorithm, set the learning rate to 0.0001, and the batch size is 8. Our module trained 250 epoch, and resize the image to  $352 \times 352$ . Finally, the initial value of deep supervision attenuation coefficient is 0.4, and every 30 epoch decays to 0.8 times the original.

Meanwhile, we adopted six common evaluation indicators to measure the performance of our network, consisting Dice, jaccard, volume overlap error (VOE), Hausdorff distance (HD), and average symmetrical surface distance (ASD).

### 4.3 Results and Analysis

**Quantitative results.** We list the quantitative results in Tables 2 and 3. MAA-Net performs the best in terms of six indicators. To verify the learning ability of our method, we compared MAA-Net with the state-of-the-art models under the same optimization, loss function and parameters settings. As observed, the mainstream res U-Net ++ [21] and sep U-Net show excellent results in liver segmentation, but for small objects, the Attention U-Net [22] outperforms sep U-Net [23]. Since it aggregated a new attention gate to assist automatically focusing on target features

with multiple shapes. From the Table 3, compared with the res U-Net ++, MAA-Net promoted by 3.8% and 4.8% on the LiTS17/3DIRCADb, respectively.

Table 4 lists the parameters, the dimension and the GFLOPs of contrast tests on the two databases. Sep U-Net imported inverse residual, yet, adding a large convolution kernel will inevitably lead to increase the parameters. Thus, MAA-Net reduces the amount of parameters to 56% compared to the same-dimensional model.

To further confirm the performance of MAA-Net, Table 5 compared with the multi-dimensional liver tumor segmentation methods on LiTS17 dataset. Specifically, compared with Triplanar FCN, the Dice of ours in liver segmentation increased by 0.20%. While in the lesion, the Dice of MAA-Net is 2.49% higher than Hy\_CompNet. When reducing the parameter amount, the effect is equivalent to some 3D models.

**Table 2** Results of liver comparative experiments on the LiTS2017 and 3DIRCADb datasets

Datasets	Model	Dice	Jaccard	VOE	HD	RVD	ASD
LiTS17	U-Net [2]	0.951	0.911	0.089	6.324	0.016	3.749
	Res U-Net ++ [21]	0.963	0.929	0.070	5.598	0.012	3.590
	Attention U-Net [22]	0.950	0.906	0.094	7.241	0.038	3.723
	sep U-Net [23]	0.956	0.916	0.084	5.261	0.037	4.231
	MAA-Net	<b>0.969</b>	<b>0.941</b>	<b>0.059</b>	<b>4.0</b>	<b>0.010</b>	<b>3.167</b>
3DIRCADb	U-Net [2]	0.943	0.892	0.108	13.92	-0.044	5.506
	Res U-Net ++ [21]	0.954	0.912	0.088	13.07	0.067	4.653
	Attention U-Net [22]	0.944	0.894	0.106	15.36	0.054	5.272
	sep U-Net [23]	0.955	0.915	0.085	14.96	0.024	4.350
	MAA-Net	<b>0.965</b>	<b>0.932</b>	<b>0.068</b>	<b>9.273</b>	<b>-0.008</b>	<b>3.894</b>

Note that the best results are marked by boldface

**Table 3** Results of tumor comparative experiments on the LiTS2017 and 3DIRCADb datasets

Datasets	Model	Dice	Jaccard	VOE	HD	RVD	ASD
LiTS17	U-Net [2]	0.613	0.634	0.366	56.25	-0.076	15.89
	Res U-Net ++ [21]	0.660	0.693	0.307	57.16	-0.073	13.63
	Attention U-Net [22]	0.639	0.621	0.379	55.90	<b>-0.067</b>	13.37
	sep U-Net [23]	0.600	0.614	0.386	52.77	-0.091	12.26
	MAA-Net	<b>0.698</b>	<b>0.694</b>	<b>0.306</b>	<b>48.31</b>	0.071	<b>10.08</b>
3DIRCADb	U-Net [2]	0.599	0.625	0.375	62.20	0.111	13.26
	Res U-Net ++ [21]	0.644	0.662	0.338	57.19	-0.075	14.60
	Attention U-Net [22]	0.627	0.653	0.347	57.42	-0.049	13.90
	sep U-Net [23]	0.608	0.648	0.352	53.17	-0.029	12.83
	MAA-Net	<b>0.692</b>	<b>0.679</b>	<b>0.321</b>	<b>50.45</b>	<b>-0.028</b>	<b>11.30</b>

Note that the best results are marked by boldface

**Table 4** Parameters, dimension and GFLOPs of comparative models

Model	Dimension	Parameters	GFLOPs
U-Net [2]	2D	8,636,802	31.03
res U-Net ++ [21]	2D	3,626,421	33.59
Attention U-Net [22]	2D	34,878,573	125.82
sep U-Net [23]	2.5D	7,376,066	19.63
MAA-Net	2.5D	<b>3,196,174</b>	<b>10.37</b>

Note that the best results are marked by boldface

**Table 5** Comparison among other mainstream method on the LiTS2017

Method	Dimension	Liver	Lesion
H-Dense U-Net [5]	3D	0.961	<b>0.722</b>
AHC-Net [4]	3D	0.965	0.690
Triplanar FCN [6]	2.5D	0.967	–
Res SegNet [17]	2.5D	0.952	0.681
Hy_CompNet [7]	2.5D	–	0.681
FSF U-Net [3]	2D	0.962	0.684
MAA-Net	2.5D	<b>0.969</b>	0.698

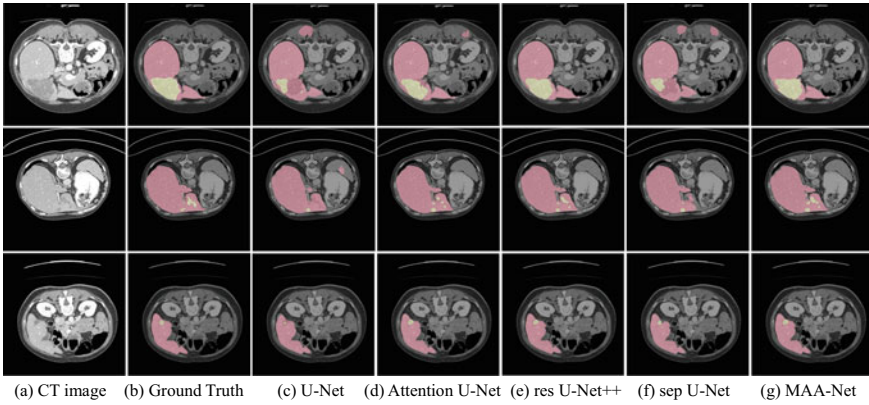
Note that the best results are marked by boldface

**Qualitative results.** We chose the representative visualizations in Fig. 4. As observed, in the first row where the liver edge contains lesions, the mainstream methods all lose the tumor area to a certain extent, and omitting the edge information in the lower right corner of the liver. The intractable issues of tumor segmentation are the high variability of tumor morphology and the indeterminacy of location. Thus, the lesion is usually under-segmentation, as the third row in Fig. 5. Overall, the consequence sufficiently illustrate the MAA-Net is a stable network with strong robustness.

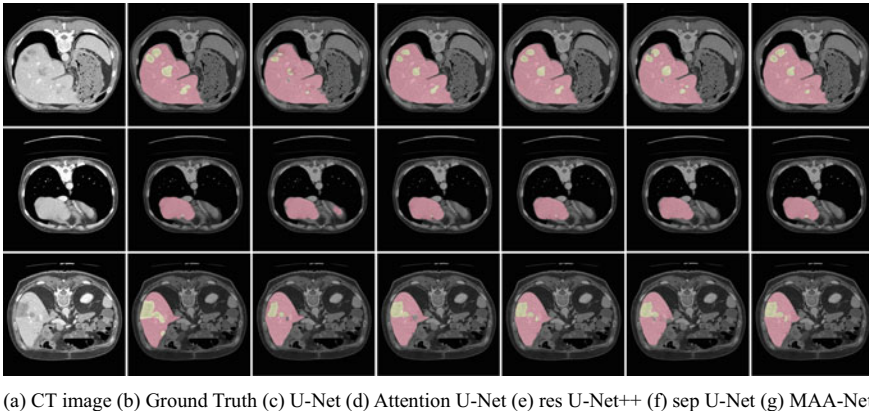
**Ablation results.** To verify the effectiveness of our MAA-Net, we evaluate the segmentation performance on 3DIRCADb. Table 6 lists the results of liver and lesion on 3DIRCADb. We consider the availability of GCA\_E unit, DFB unit and GCA unit. We found the scores of each module added on U-Net increased. Specially, the structure with hierarchical encoder-decoder and bidirectional feature fusion units has achieved optimal results in multiple indicators. Among them, the Dice of MAA-Net advanced by 8.5% and 8.3%, respectively. Figure 6 visualizes the ablation experimental results, where the red area is the liver and the yellow area is tumor. For the convenience of observation, the test images in the figure are all preprocessed.

Obviously, all models reached excellent results in liver segmentation and is corresponding to the quantitative results in Table 6. There is a certain degree of under or over segmentation to varying extent. On the contrast, MAA-Net outperforms other methods when addressing with borderline lesions. Similar to the first row in Fig. 6, our algorithm successfully segmented two tiny tumors and closer to the ground truth.





**Fig. 4** Visualization of comparative experiments on the LiTS2017 database (a original images; b labels; c U-Net; d Attention U-Net; e res U-Net ++; f sep U-Net; g ours)

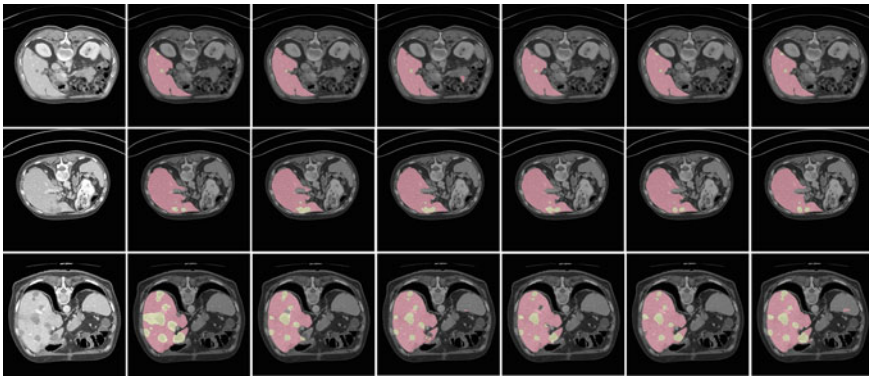


**Fig. 5** Visualization of comparative experiments on the 3DIRCADb database (a original images; b labels; c U-Net; d Attention U-Net; e res U-Net ++; f sep U-Net; g ours)

**Table 6** Results of tumor ablation experiments on the LiTS2017 and 3DIRCADb datasets

Datasets	Model	Dice	Jaccard	VOE	HD	RVD	ASD
LiTS17	U-Net	0.613	0.634	0.366	56.25	-0.076	15.89
	+GCA	0.653	0.671	0.329	52.38	0.087	13.86
	+GCA_E	0.673	0.657	0.343	51.10	-0.080	<b>8.216</b>
	+DFB	0.680	0.659	0.341	49.34	<b>0.047</b>	8.245
	MAA-Net	<b>0.698</b>	<b>0.664</b>	<b>0.336</b>	<b>48.31</b>	0.071	9.078
3DIRCADb	U-Net	0.599	0.625	0.375	62.20	0.111	13.26
	+GCA	0.603	0.582	0.418	61.63	-0.087	12.86
	+GCA_E	0.642	0.673	0.327	60.21	-0.034	12.73
	+DFB	0.658	0.674	0.326	56.64	0.050	11.93
	MAA-Net	<b>0.682</b>	<b>0.671</b>	<b>0.329</b>	<b>50.45</b>	<b>-0.028</b>	<b>11.30</b>

Note that the best results are marked by boldface



(a) CT image (b) Ground Truth (c) U-Net (d) +GCA (e) +GCA\_E (f) +DFB (g) MAA-Net

**Fig. 6** Visualization of ablation experiments on the 3DIRCADb database (a original images; b labels; c U-Net; d +GCA; e +GCA\_E; f +DFB; g ours)

## 5 Conclusion

In this paper, we have proposed a 2.5D lightweight structure for segmenting liver and lesion in abdominal CT scans. Our framework includes a multi-scale encoder-decoder framework MAA-Net, which is based on the global context aware and bidirectional feature fusion. In addition, we trained our model using the novel form of input, and it realized the information interaction between different levels. Extensive experiments on the public two databases showed our MAA-Net can reduce the amount of calculations by 57.5% when the evaluation indicators are not much different. The results also declared the compromise between computation resources and segmentation accuracy. However, the algorithm has certain shortcomings when handing the

adhering lesion issue. Hence, our future work would concentrate on the multi-view fusion, adopting sagittal, coronal and axial orthogonal planes to achieve other 2.5D strategy for multiple lesion detection.

## References

1. Long J., Shelhamer E., Darrell T.: Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(4), 640–651 (2015).
2. Ronneberger O., Fischer P., Brox T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. Springer, Cham, (2015).
3. Qiao W C., Hunag M.: Feature selection and residual fusion segmentation network for liver tumor. *Journal of Image and Graphics*, 27(03), 838–849 (2022). (in Chinese)
4. Jiang H., Shi T., Bai Z.: AHC-Net: An Application of Attention Mechanism and Hybrid Connection for Liver Tumor Segmentation in CT Volumes. *IEEE Access*, 7, pp. 24898–24909 (2019). <https://doi.org/10.1109/ACCESS.2019.2899608>
5. Li X., Chen H., Qi X.: H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes. *IEEE Transactions on Medical Imaging*, (2018). <https://doi.org/https://doi.org/10.1109/tmi.2018.2845918>
6. Wang Z.: Triplanar Convolutional Neural Network for Automatic Liver and Tumor Image Segmentation. *International Journal of Performability Engineering* 14(12) (2018). <https://doi.org/10.23940/ijpe.18.12.p24.31513158>
7. Dey R, Hong Y, Hybrid Cascaded Neural Network for Liver Lesion Segmentation. *International Symposium on Biomedical Imaging. IEEE* (2020). <https://doi.org/https://doi.org/10.1109/ISB145749.2020.9098656>
8. Szegedy C., Vanhoucke V., Ioffe S.: Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016).
9. Guo N., Bai Z Y.: The integration of attention mechanism and dense atrous convolution for lung image segmentation. *Journal of Image and Graphics*, 26(09): 2146–2155(2021). (in Chinese)
10. Lv Peiqing, Wang Jinke, Wang Haiying. 2.5D lightweight RIU-Net for automatic liver and tumor segmentation from CT. *Biomedical Signal Processing and Control*, 75 (2022). <https://doi.org/10.1016/j.bspc.2022.103567>
11. He K M., Zhang X Y., Ren S Q and Sun J.: 2016. Deep residual learning for image recognition // *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770–778* (2016). [DOI: 10. 1109 / CVPR. 2016. 90]
12. Jie, Shen, Samuel, et al. Squeeze-and-Excitation Networks. *IEEE transactions on pattern analysis and machine intelligence*, (2019). <https://doi.org/10.1109/tpami.2019.2913372>
13. Woo S , Park J , Lee J Y , et al. CBAM: Convolutional Block Attention Module. Springer, Cham, (2018).
14. Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, (2020). <https://doi.org/10.1109/CVPR42600.2020.01155>
15. Li S D, Bai Z Y. Multiorgan lesion detection and segmentation based on deep learning. *Journal of Image and Graphics*, 26(11):2723–2731 (2021). (in Chinese)
16. Ji A., Bo D B., Shuai W C.: COVID-19 Lung Infection Segmentation with A Novel Two-Stage Cross-Domain Transfer Learning Framework. *Medical Image Analysis*, (2021). <https://doi.org/10.1016/j.media.2021.102205>
17. Wardhana G., Naghibi H., Sirmacek B.: Toward reliable automatic liver and tumor segmentation using convolutional neural network based on 2.5D models. *International Journal of Computer Assisted Radiology and Surgery*, 16(12), (2020). <https://doi.org/10.1007/s11548-020-02292-y>

18. Hou Q., Zhou D., Feng J.: Coordinate Attention for Efficient Mobile Network Design, (2021). <https://doi.org/10.1109/CVPR46437.2021.01350>
19. Bilic P., Christ P F., Vorontsov E.: The Liver Tumor Segmentation Benchmark (LiTS), (2019). <https://doi.org/10.48550/arXiv.1901.04056>
20. Soler L., Hostettler A., Agnus V.: 3D image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database. IRCAD, Strasbourg, France, Tech. Rep, (2010).
21. Jha D., Smedsrud P H., Riegler M A.: ResUNet++: An Advanced Architecture for Medical Image Segmentation. 21st IEEE International Symposium on Multimedia. IEEE, (2019). <https://doi.org/10.1109/ISM46123.2019.00049>
22. Oktay O., Schlemper J., Folgoc L L.: Attention U-Net: Learning Where to Look for the Pancreas. (2018). <https://doi.org/10.48550/arXiv.1804.03999>
23. Lei W., Mei H, Sun Z.: Automatic Segmentation of Organs-at-Risk from Head-and-Neck CT using Separable Convolutional Neural Network with Hard-Region-Weighted Loss. (2021). <https://doi.org/10.1016/j.neucom.2021.01.135>

# Registration of Medical Image Sequences Using Auto-differentiation



Tomas Vicar, Roman Jakubicek, Jiri Chmelik, and Radim Kolar

**Abstract** This paper focuses on image registration using the automatic differentiation of deep learning frameworks. Specifically, a method for the registration of image sequences is proposed and tested on retinal video ophthalmoscopic data and brain DCE MR images. PyTorch auto-differentiation has been used as a core of an optimisation tool to find the optimal image transformation parameters. It allows us to easily design a loss function for our registration tasks. The image registration was achieved by simultaneous registration of all images using a global loss function without the need of the reference frame.

**Keywords** Medical image registration · Auto-differentiation · Deep learning frameworks · Gradient-based optimisation · Video stabilisation

## 1 Introduction

Image registration is a process that leads to a geometrical alignment of images acquired at different times, from different points of view, and/or by different sensors [18]. This is among the most important image processing tasks and continues to be an active research topic with several important applications in medical imaging. Image registration is defined as the search for a set of transformation parameters  $\mathbf{u}$  of a spatial transformation  $\mathbf{T}_{\mathbf{u}}(\mathbf{x})$  that is capable of aligning the fixed image  $f_F$  with the moving image  $f_M$  as [17]:

---

T. Vicar · R. Jakubicek (✉) · J. Chmelik · R. Kolar  
Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication,  
Brno University of Technology, Brno, Czech Republic  
e-mail: jakubicek@vut.cz

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
R. Su et al. (eds.), *Medical Imaging and Computer-Aided Diagnosis*, Lecture Notes  
in Electrical Engineering 810, [https://doi.org/10.1007/978-981-16-6775-6\\_15](https://doi.org/10.1007/978-981-16-6775-6_15)

169

$$\mathbf{u}_{\text{opt}} = \arg \min_{\mathbf{u}} \left[ L \left( f_F(\mathbf{x}), f_M(\mathbf{T}_{\mathbf{u}}(\mathbf{x})) \right) + \lambda R(\mathbf{u}) \right], \quad (1)$$

where  $L$  is the loss function designed to have a minimum when the two images are aligned, and  $R(\mathbf{u})$  is a regularisation term. In this paper, we are using the affine transformations defined as

$$\mathbf{T}_{\mathbf{u}}(\mathbf{x}) := \mathbf{A}\mathbf{x}, \quad (2)$$

where  $\mathbf{A}$  is the linear transformation matrix of size  $3 \times 3$  for 2D and  $4 \times 4$  for 3D, and  $\mathbf{X}$  are homogeneous coordinates.

Although deep learning-based methods outperformed classical methods in most computer vision tasks, their application to image registration is still very limited [16]. The reason is that the main success of deep learning lies in supervised applications with the availability of large labelled datasets [1], which is usually not the case for image registration problems. However, deep learning frameworks can serve as an easy-to-use gradient-based optimisation tool, thanks to available auto-differentiation and implementation of well-performing optimisers and other useful tools. Moreover, they can be easily extended from 2D to 3D and provide access to GPU computation, which can significantly speed up the calculations.

Currently, several image registration frameworks are available (see [2]), but the most popular tool for medical image registration is Elastix [5, 14], which is a part of the Insight Segmentation and Registration Toolkit framework [9]. Elastix is based on gradient optimisation, providing various registration options including rigid, affine, and nonrigid transformation models, monomodal, and multimodal loss functions, etc. Moreover, various wrappers introduce this tool to other programming languages like Python or Java. It is written in C++ including GPU implementations; however, it is not easy to include custom modifications, mainly because the gradient calculation needs to be defined; thus, it is not suitable for rapid prototyping.

The proposed approach uses PyTorch capabilities [11] that enable the fast and simple development of a new registration approach. Therefore, this approach benefits from the active development of this deep learning framework and the vibrant community around it. Similar efforts have been made by the Kornia authors [12], which uses an easy-to-use Pytorch-based registration API with access to basic image transformations (affine) and losses for the registration of two monomodal images including the pyramidal approach. In AirLab [13] the authors developed a PyTorch-based tool for the rapid development and testing of registration methods, where users can introduce custom losses, transformations, or regularisations; however, neither allows more complex adjustments as we require.

Here, we focus on the application of this approach to a specific task, which is stabilisation of image sequence (registration of multiple images together). This problem can be approached by selecting a reference frame, where other frames are individually aligned to this frame [6]; however, it leads to an increase in registration errors with increasing differences from the reference frame (especially in sequences with significant information changes). In order to overcome these problems, we have

defined a global loss function, which leads to the registration of all frames together without the need for a specific reference frame.

## 2 Methods

### 2.1 Image Sequence Registration

The problem of image sequence registration can be formally defined as

$$\arg \min_{\mathbf{u}_1, \dots, \mathbf{u}_n} L\left(f_1(\mathbf{T}_{\mathbf{u}_1}(\mathbf{x})), w_1(\mathbf{T}_{\mathbf{u}_1}(\mathbf{x})), \dots, f_n(\mathbf{T}_{\mathbf{u}_n}(\mathbf{x})), w_n(\mathbf{T}_{\mathbf{u}_n}(\mathbf{x}))\right), \quad (3)$$

where we search for parameters  $\mathbf{u}_i$  of individual transformations  $\mathbf{T}_{\mathbf{u}_i}(\mathbf{x})$  for images  $f_i$  in the sequence. It follows that there is no reference image, but all images are moving images. Furthermore, image masks (positional weights)  $w_i$  are very important for successful registration [3], where they are mainly required to ignore the image border, which can significantly bias the loss. Masks can be binary or in the form of weights, and they can be more important to a specific part of the image. The weights are transformed together with each image. We are interested in the loss function, which will ensure that all images are aligned; therefore, variance over time can be a suitable loss:

$$\begin{aligned} & L_k\left(f_1(\mathbf{T}_{\mathbf{u}_1}(\mathbf{x}_k)), w_1(\mathbf{T}_{\mathbf{u}_1}(\mathbf{x}_k)), \dots, f_n(\mathbf{T}_{\mathbf{u}_n}(\mathbf{x}_k)), w_n(\mathbf{T}_{\mathbf{u}_n}(\mathbf{x}_k))\right) \\ &= \frac{1}{N} \sum_i \left( w_i(\mathbf{T}_{\mathbf{u}_i}(\mathbf{x}_k)) f_i(\mathbf{T}_{\mathbf{u}_i}(\mathbf{x}_k)) - \frac{1}{N} \sum_j w_j(\mathbf{T}_{\mathbf{u}_j}(\mathbf{x}_k)) f_j(\mathbf{T}_{\mathbf{u}_j}(\mathbf{x}_k)) \right)^2, \quad (4) \end{aligned}$$

where each image  $f_i$  is weighted by its mask  $w_i$ ; the final loss function is the mean variance of the individual positions  $k$ . However, we need to restrict this sum to valid positions, where the majority of masks are non-zero. For this reason, the final loss is calculated as  $L = \frac{1}{|V|} \sum_{k \in V} L_k$ , where  $V$  is the set of valid positions and  $|V|$  is the number of positions in this set, where valid positions are the positions where at least 50 % of transformed masks are nonzero.

This approach does not require reference image; however, it is invariant to the same additional transformation applied to all images. For this reason, we need to introduce regularisation, which will keep the 'average transformation' close to identity. It can be done by this regularisation term:

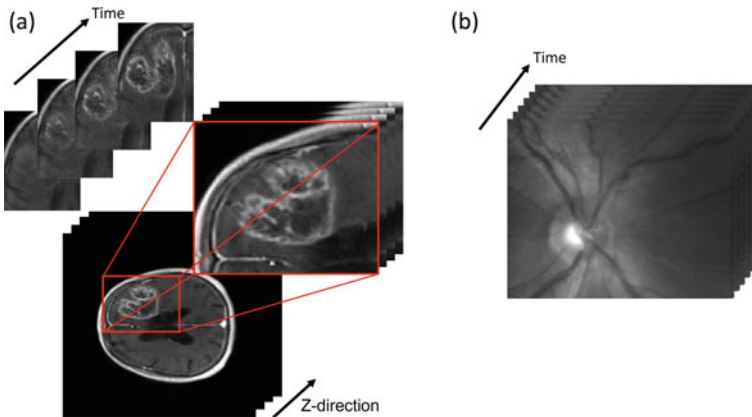
$$R_{identity}(\mathbf{u}_1, \dots, \mathbf{u}_n) = \left\| I - \frac{1}{N} \sum_{i=1}^N A_i \right\|_2^2, \quad (5)$$

where  $I$  is the identity matrix,  $A_i$  is the affine transformation matrix for the transformation  $T_{u_i}(\cdot)$  and  $\|\cdot\|_2^2$  is the  $L_2$ -norm squared.

## 2.2 Experimental Data

Retinal image sequences were acquired using a previously developed video ophthalmoscope (VO) [15], recently modified to a multispectral monocular version [7]. This device acquires retinal video of the optic nerve head and the peripapillary area with a field of view of  $20^\circ \times 17^\circ$  ( $1224 \times 970$  pixels). This corresponds to an approximate area of  $6 \times 5$  mm in the retina for a subject with an axial length of 24 mm. The frame rate of the CMOS camera was set at 25 frames per second (exposure time approximately 40 ms). The light power in the plane of the eye pupil was less than  $15 \mu\text{W}$ . The length of the sequences is 5 s. There are two main artefacts that can distort individual frames and cause difficulties in the registration process. The first artefact is caused by the limited exposure time (40 ms), which can subsequently cause motion blur due to involuntary eye movements during acquisition. The second major artefact can occur during eye blinking, when a strong reflection from the eyelid causes saturation of the image intensity. An example of a VO retinal image sequence is shown in Fig. 1b.

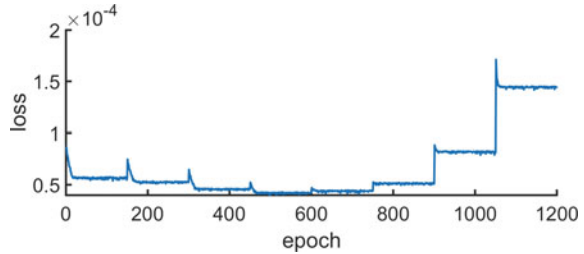
The second tested dataset contains volumetric brain magnetic resonance (MR) images of oncological patients with glioma acquired at St. Anne's University Hospital Brno. For scanning, the GE Discovery MR750 3.0 T with a voxel size of  $0.937 \times 0.937 \text{ mm}^2$  with 6 mm of slice thickness. During examination, a contrast agent is injected and scanned over time (60 time points in total); therefore, the output volumes are T1-weighted dynamic contrast enhanced (DCE) MR data sized as  $240 \times 240 \times$



**Fig. 1** Example of images used for testing of proposed registration method. **a** MR dynamic contrast enhanced data, 3D time-lapse sequences, registration was applied to cropped the part of the images with the tumour. **b** Video-ophthalmoscopic retinal image sequence data



**Fig. 2** Example of loss function development during registration for VO sequence. Steps are caused by the resolution change of the pyramidal approach of registration



150 × 60 voxels. An example of DCE MR data is shown in Fig. 1a, where registration was applied on images cropped only to the segmented tumour region, where this registration is necessary for its subsequent analysis.

### 2.3 Implementation Details

Due to GPU memory limitations, it is problematic to optimise transformations of all images in the sequence simultaneously in a single optimisation step; however, similarly to deep learning methods, the mini-batch processing approach [8] can be used.

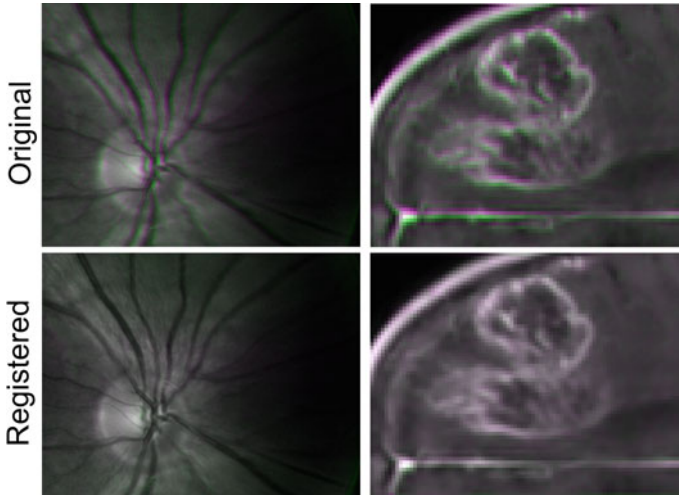
For VO images, which contain larger movement, the pyramidal approach was used, where the images were first registered with a smaller resolution and then their registration was refined with gradually increasing resolution. Specifically, the images were subsampled by factors of  $\sqrt{64}$ ,  $\sqrt{32}$ ,  $\sqrt{16}$ ,  $\sqrt{8}$ ,  $\sqrt{4}$ ,  $\sqrt{2}$  and 1. Transformation parameters were optimised using Adam optimiser [4] with the parameter of 1st and 2nd moment estimates set to 0.9 and 0.999, respectively. The learning rate 0.0002 (0.002 without pyramidal approach) decayed to 80% after 120, 140 and 150 epochs were used. The regularisation factor of  $10^{-5}$  was used. An example of a loss function that includes steps caused by the pyramidal approach is shown in Fig. 2.

Movements in the MR data are not as frequent and significant (in a range of units of voxels) as in the VO images, so it was not necessary to use the pyramidal approach. However, the learning rate settings, including decay and number of epochs, were the same as in the case of VO images. However, the source code had to be extended by editing transformation matrices and geometric transforms for 3D images, modifying the calculation of the loss function.

The code is publicly available on the [GitHub repository](#).

## 3 Results and Discussion

The proposed method was successfully applied to both VO retinal sequences and DCE MR images. Examples of unregistered and randomly selected registered frames

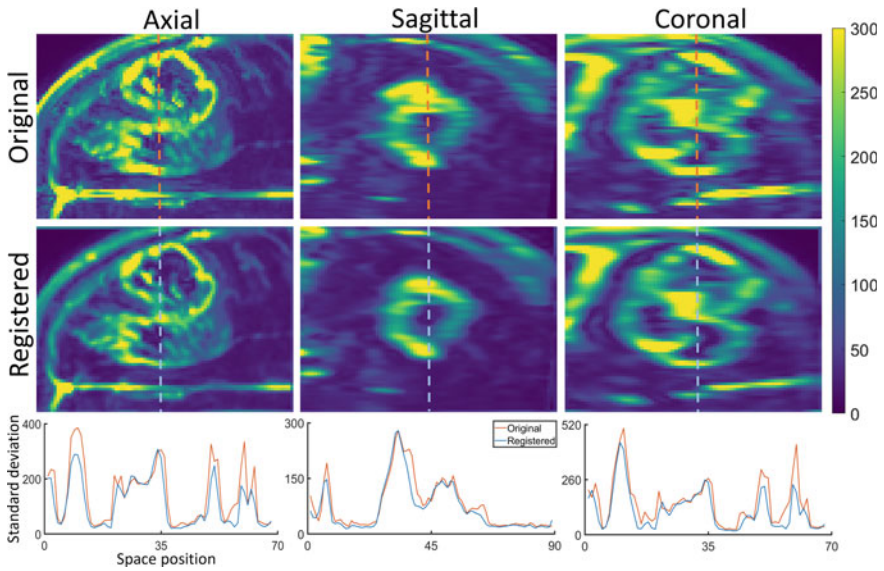


**Fig. 3** Visualisation of the differences between two sample frames in the original sequence and in the registered sequence. The images are composites (overlays) of two images, where one is shown in cyan and the other in magenta

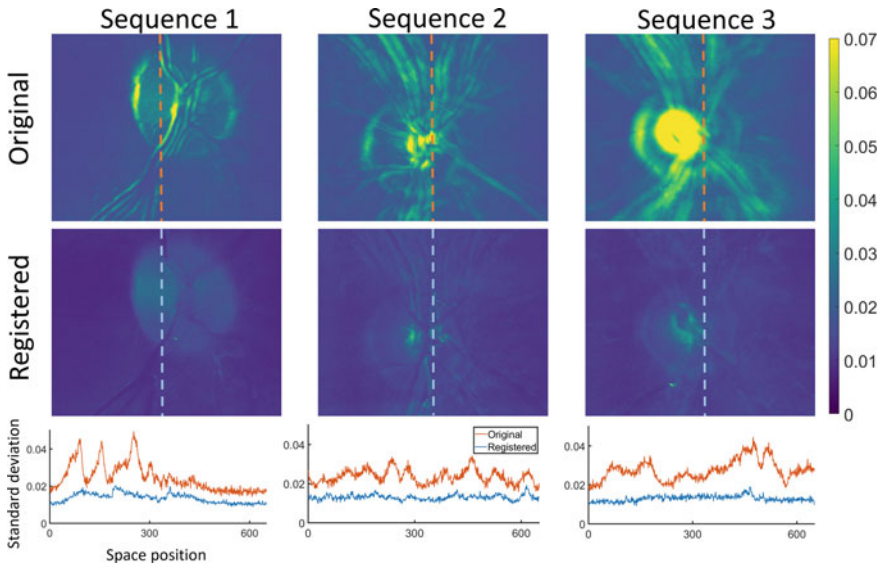
from the sequence are visualised as colour composites in Fig. 3, where the registered frames show significantly better overlap.

Examples of registration quality are also visualised with the standard deviation values over time in Figs. 4 and 5 for VO and MR images, respectively. As can be seen in both figures, the original images contain significantly more movements between frames; thus, there is a more significant improvement in the standard deviation caused by registration.

We have evaluated the variance (equivalent to our loss without regularisation) for five VO sequences for various data to compare the quality of registration for those settings in Table 1. Here, the variance of the original data compared to the data registered decreased significantly from  $6.07 \times 10^{-4}$  to  $1.45 \times 10^{-4}$ . We have also compared the proposed approach with a fixed learning rate of 0.0002 with the results, where the learning rate was optimised for each image individually. It provided a decrease in the variance from  $1.45 \times 10^{-4}$  to  $1.28 \times 10^{-4}$ , however, this requires 15 registration evaluations using Bayesian optimisation [10]. Next, we have found that there is a significant effect of the pyramidal approach, where registration without the pyramidal approach achieved only variance  $1.84 \times 10^{-4}$ . We have also tested registration without the mini-batch approach (with a single mini-batch on a GPU with sufficient memory) and reached a very similar result to the technique with the mini-batch (fixed lr.) of variance  $1.43 \times 10^{-4}$ ; however, application of the mini-batch is necessary if the GPU memory is not sufficient to register the whole sequence at once. Registration without regularisation achieved the same numerical result as regularised one; however, registration without this regularisation can result in the movement of the entire sequence, as can be seen in Fig. 6. We have also evaluated



**Fig. 4** Visualisation of standard deviation in-time for original and registered MR DCE images. The bottom part shows the value profile through the images in the upper part defined by the red and blue lines. Lower standard deviation values after registration (blue curve) shows significant improvement in sequence stability in time



**Fig. 5** Visualisation of standard deviation in-time for original and registered VO sequences. The bottom part shows slices through the images at the top. Lower standard deviation after registration shows significant improvement in sequence stability in time

**Table 1** Resulting variance of various registration settings for VO data

Seq.	Variance (opt. lr.)	Variance (fixed lr.)	w/o pyramid	w/o mini-batch	w/o regularisation	Register to 1st frame	Original
1	1.19	1.60	1.76	1.62	1.60	1.48	3.46
2	1.48	1.49	2.10	1.51	1.49	1.59	4.87
3	1.43	1.59	2.06	1.70	1.60	1.87	8.20
4	1.62	1.89	2.04	1.67	1.88	1.86	9.08
5	0.66	0.67	1.22	0.67	0.66	0.69	4.73
Avg.	1.28	1.45	1.84	1.43	1.45	1.50	6.07

Values are calculated for image values in the range 0–1 and values are  $10^{-4} \times$  variance; opt. lr. is optimal learning rate different for every image (fixed lr. is used otherwise), w/o pyramid is without pyramidal approach (original scale is used), w/o mini-batch is with all data in a single batch

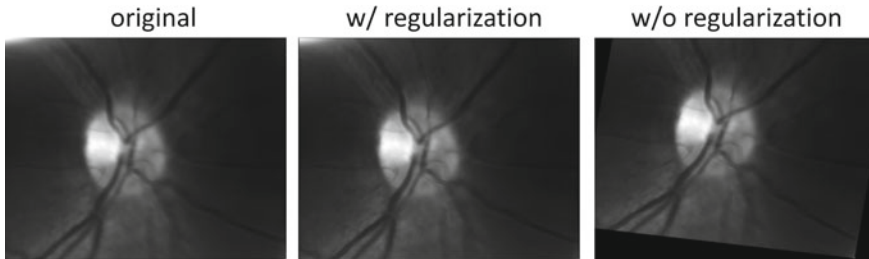
the approach of standard registration to the first frame of the sequence, which was calculated using our same PyTorch implementation, where we did not transform the first image and calculate the loss as the sum of mean squared errors between the first frame and other individual frames. The results of the registration in the first frame achieved slightly worse results of variance  $1.50 \times 10^{-4}$ .

In the case of the DCE MR dataset, the proposed algorithm provides a decrease in the overall standard deviation value from 86.72 to 73.98 after registration for a selected scan (Fig. 4). There was a reduction in the standard deviation values of 7.68 on average (evaluated in 3 samples). However, the problem is the nonzero residual standard deviation caused by the change in image contrast over time due to the contrast agent; therefore, each scan has a different value of standard deviation.

Similar effects to the VO listed in Table 1 were also achieved for the DCE MR data, where regularisation protects the optimisation process from significantly shifting the 3D data in the Z direction out of the image range that is replaced by zeros. This significantly but incorrectly increases the value of the loss function and leads to incorrect registration results. Thus, during optimisation, it was unnecessary to regularise the translation in all directions.

The effect of cropping the MR data resulted in a significant reduction in memory and computational requirements, naturally leading to faster convergence but ultimately to the same solution. However, the use of mini-batches with significantly longer computational time and slower convergence was necessary when registering uncropped images. The computational time of the proposed approach was 10, 8, and 0.5 min for VO, uncropped MRI, and cropped MRI, respectively (on Intel Core i9-10900KF with NVIDIA GeForce RTX3090).

The registration of all frames together with the proposed loss function has other advantages; besides the slightly better result, the reference frame can be corrupted or very different from other frames. Moreover, it can influence the results, for example, it can register images in different phases of the cardiac cycle compared to the reference frame differently.



**Fig. 6** Example of sequence movement if regularisation is not used. Time-averages of the original image sequence, the sequence with regularisation and the sequence without regularisation. Non-regularised registration causes the rotation of the entire registered sequence

Similarly to this application, the PyTorch implementation of registration provides the perfect tool for customised gradient-based registration approaches. In addition to extensions of multimodal registration with mutual information [17], it can be used to define task-specific loss functions and transformations. The major problem with gradient-based optimisation is that it converges to local optima only, making it very sensitive to the setting of hyperparameters. This can be partially overcome by wrapping the whole registration process into hyperparameter optimisation; this is tractable due to the utilisation of GPU, which significantly reduces computation time, where we have already tested optimisation of the learning rate.

## 4 Conclusions

This paper focuses on the registration of medical image sequences using the PyTorch implementation of gradient-descent-based registration. Implementation of the specialised loss function for the registration of image sequences was tested on both 2D+time video ophthalmoscope retinal images and 3D+time MR based dynamic contrast-enhanced images, where in both cases we have achieved sufficient registration quality for further processing. This shows that our implementation of image registration using a deep learning framework is very suitable for the rapid development of new registration approaches.

**Acknowledgements** This work is supported by the Czech Science Foundation project no. 18-24089S. Computational resources were supplied by the project “e-Infrastruktura CZ” (e-INFRA LM2018140) provided within the programme Projects of Large Research, Development, and Innovation Infrastructures. The authors also acknowledge the contribution of St. Anne’s University Hospital in Brno, which provided MRI data.

## References

1. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. *Machine Vision and Applications* 31(1), 1–18 (2020)
2. Keszei, A.P., Berkels, B., Deserno, T.M.: Survey of non-rigid registration tools in medicine. *Journal of digital imaging* 30(1), 102–116 (2017)
3. Ketcha, M.D., De Silva, T., Uneri, A., Kleinszig, G., Vogt, S., Wolinsky, J.P., Siewerdsen, J.H.: Automatic masking for robust 3d-2d image registration in image-guided spine surgery. In: *Medical Imaging 2016: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 9786, pp. 98–104. SPIE (2016)
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980)* (2014)
5. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging* 29(1), 196–205 (2009)
6. Kolar, R., Tornow, R., Odstrcilik, J., Liberдова, I., et al.: Registration of retinal sequences from new video-ophthalmoscopic camera. *Biomedical engineering online* 15(1), 1–17 (2016)
7. Kolar, R., Vicar, T., Odstrcilik, J., Valterova, E., Skorkovska, K., Kralik, M., Tornow, R.P.: Multispectral retinal video-ophthalmoscope with fiber optic illumination. *Journal of Biophotonics* 15(9), e202200094 (2022)
8. Li, M., Zhang, T., Chen, Y., Smola, A.J.: Efficient mini-batch training for stochastic optimization. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 661–670 (2014)
9. McCormick, M., Liu, X., Ibanez, L., Jomier, J., Marion, C.: Itk: enabling reproducible research and open science. *Frontiers in Neuroinformatics* 8 (2014), <https://www.frontiersin.org/articles/10.3389/fninf.2014.00013>
10. Nogueira, F.: Bayesian Optimization: Open source constrained global optimization tool for Python (2014), <https://github.com/fmfn/BayesianOptimization>
11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimselshin, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems* 32, pp. 8024–8035. Curran Associates, Inc. (2019), <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
12. Riba, E., Mishkin, D., Ponsa, D., Rublee, E., Bradski, G.: Kornia: an open source differentiable computer vision library for pytorch. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3674–3683 (2020)
13. Sandkühler, R., Jud, C., Andermatt, S., Cattin, P.C.: Airlab: autograd image registration laboratory. *arXiv preprint [arXiv:1806.09907](https://arxiv.org/abs/1806.09907)* (2018)
14. Shamonin, D.P., Bron, E.E., Lelieveldt, B.P., Smits, M., Klein, S., Staring, M., Initiative, A.D.N.: Fast parallel image registration on cpu and gpu for diagnostic classification of alzheimer’s disease. *Frontiers in neuroinformatics* 7, 50 (2014)
15. Tornow, R.P., Odstrcilik, J., Kolar, R.: Time-resolved quantitative inter-eye comparison of cardiac cycle-induced blood volume changes in the human retina. *Biomed. Opt. Express* 9(12), 6237–6254 (2018)
16. Xiao, H., Teng, X., Liu, C., Li, T., Ren, G., Yang, R., Shen, D., Cai, J.: A review of deep learning-based three-dimensional medical image registration methods. *Quantitative Imaging in Medicine and Surgery* 11(12), 4895 (2021)
17. Xu, R., Chen, Y.W., Tang, S.Y., Morikawa, S., Kurumi, Y.: Parzen-window based normalized mutual information for medical image registration. *IEICE transactions on information and systems* 91(1), 132–144 (2008)
18. Zitova, B., Flusser, J.: Image registration methods: a survey. *Image and vision computing* 21(11), 977–1000 (2003)

# Small Animal Imaging: Iterative Algorithms Combined with Regularization Schemes, an Application to a Dual-Head Small Animal PET



**Evangelia Karali**

**Abstract** Iterative algorithms have nowadays gained an enormous clinical interest and applications, usually in their ordered subsets versions accompanied with regularization approaches. In this study the ordered subset version of expectation maximization maximum likelihood algorithm (OSEM) combined with two different regularization techniques is evaluated towards reconstruction image resolution and image quality. The regularization methods are median-root prior (MRP) and Total Variation (TV). Different regularization mask schemes and different regularization multiplicative factors are compared. The evaluation study is based on image reconstruction contrast to noise ratios (CNRs), reconstruction time and final image resolution. Simulation data of a derenzo-like phantom, taking into account a rotated camera of a standard small animal PET system, is used. Results show that reconstruction methods combined with MRP gives a better image quality for all sized objects. The multiplicative factor in MRP is small while in TV can be over 0.1. Image resolution is a function of reconstruction approach.

**Keywords** Small animal imaging · OSEM · Regularization schemes · Median-root prior · Total variation

## 1 Introduction

Small animal imaging is closely related to clinical application, because small animal PET systems evaluate radiopharmaceuticals distribution and labelling, clinical systems technology and architecture, clinical protocols and reconstruction algorithms. Such systems are capable of an image resolution below 1 mm at FOV's (field of view) edges and high radiation sensitivity [1].

Iterative reconstruction schemes are now mandatory in every clinical task. Analytical methods have been thorough studied the previous decades and have been proved inferior as far image quality concerns. Iterative algorithms are divided to:

---

E. Karali (✉)

University of West Attica, Ag Spyridonos St, P.O. box 12243, Egaleo, Athens, Greece  
e-mail: [ekarali@uniwa.gr](mailto:ekarali@uniwa.gr)

- algebraic techniques where a system of equations, less in number than the unknown variables, is tried to be solved through a successive iterative procedure
- statistical methods that model all the physical phenomena during acquisition process under known statistical distributions. This set of algorithms is further divided to Poisson-like approaches (like EM-ML, Expectation Maximization Maximum Likelihood algorithm) and Gaussian schemes (like ISRA-Image Space Reconstruction Algorithm, WLS-Weighted Least Squares methods) [1].

The aforementioned reconstruction schemes are usually applied in:

- simultaneous ordered subsets (OS) versions like OSEM [2] (OS-Expectation Maximization or the OS version of EM-ML), where a subset of collected data is reconstructed at every subiteration, next subset starts from the resulted image estimate of the previous subset, these methods are famous for their reconstruction speed.
- or Row action ordered subsets versions like OS-RAMLA (OS-Row Action Maximum Likelihood Algorithm), where all the pixels that intersect a specific line of response (LOR) are reconstructed each time.

They usually combined with relaxation parameters to speed up the process and regularization approaches, where a priori information about the subject under study is used. The most significant priors are Median-Root-Prior and Total Variation Prior [3–7].

In this study OSEM is used accompanied with two regularization schemes, namely OSEM-MRP and OSEM-TV. Such a study has not been conducted previously. Regularization approaches are applied after every subiteration of OSEM. This paper presents a study from a prototype small animal PET using a derenzo-like phantom and introduce iterative schemes where the prior is applied during subiterations, without post-processing. Iterative schemes shown here are applied to 3D sinograms followed by a SSRB (single slice rebinning) method that produces 2D sinograms.

## 2 Theory

Suppose  $y$  are the collected data (here in sinogram mode, a matrix where photons in every angle for every line of response-LOR are shown) and  $x$  image data. Collected data and image data are linearly connected, according to equation:

$$y = A^T x \quad (1)$$

where  $A$  is the System or Probability Matrix, a matrix variable that models all the physical phenomena during data acquisition process, namely positron range, photons scatter and attenuation, photons acollinearity, and scanner geometrical characteristics (number of rings, angle of rotation, number of pixels in a block detector, number of block detectors, pixel size, image size). Element  $a_{ij}$  of matrix  $A$  represents the



probability a pair of photons travelling from image pixel  $i$  to be detected from the two antidiametrical detector pixels that define LOR  $j$ . Because  $A$  is not quadratic,  $A^{-1}$  cannot be calculated so Eq. (1) cannot be direct solved, to find image  $x$ .

Statistical Iterative formulas give the best solution to (1), in conjunction with the statistical phenomenon they try to model. OSEM is based on the assumption that collected data follow a Poisson distribution with mean value  $\sum_{i=1}^N a_{ij}x_i$ , where  $N$  the total image pixels number. The iterative step of OSEM in  $k$ th iteration for subset  $n$  is:

$$\text{OSEM(or OS-EM-ML)} : x_i^k = x_i^{k-1} \sum_{j \in S_n} \frac{a_{ij}y_j}{\left(\sum_{i'=1}^N a_{i'j}x_{i'}^{k-1}\right)} \tag{2}$$

When iterative methods are combined with regularization approaches, usually an one step late iterative scheme is chosen, where the prior is applied on the previous image pixel value [6]. However, there are techniques where the prior is a part of image pixel update and is taking into account for extracting current pixel's value [7]. Here an one step late method is considered.

MRP prior is among the most famous. This penalty function is actually a Gaussian distribution where the median value in the vicinity of pixel  $i$  is the distribution's mean value. Assuming an area  $m \times m$  around pixel  $i$ , the median value of the  $m \times m$  matrix is calculated and applied in every pixel update step:

$$\text{OSEM-MRP} : x_i^{k-1} = \frac{x_i^{k-1}}{1 + b \frac{x_i^{k-1} - \text{med}(x_i^{k-1}, m)}{\text{med}(x_i^{k-1}, m)}} \sum_{j \in S_n} \frac{a_{ij}y_j}{\left(\sum_{i'=1}^N a_{i'j}x_{i'}^{k-1}\right)} \tag{3}$$

where  $b < 1$  a multiplicative factor.  $b$  cannot take a big value because image smoothing increases noise. With a small  $b$ -value image edges are preserved.

Total Variation prior [8] takes into account the sum of pixel value differences in the vicinity of pixel  $i$ , so OSEM-TV update step is:

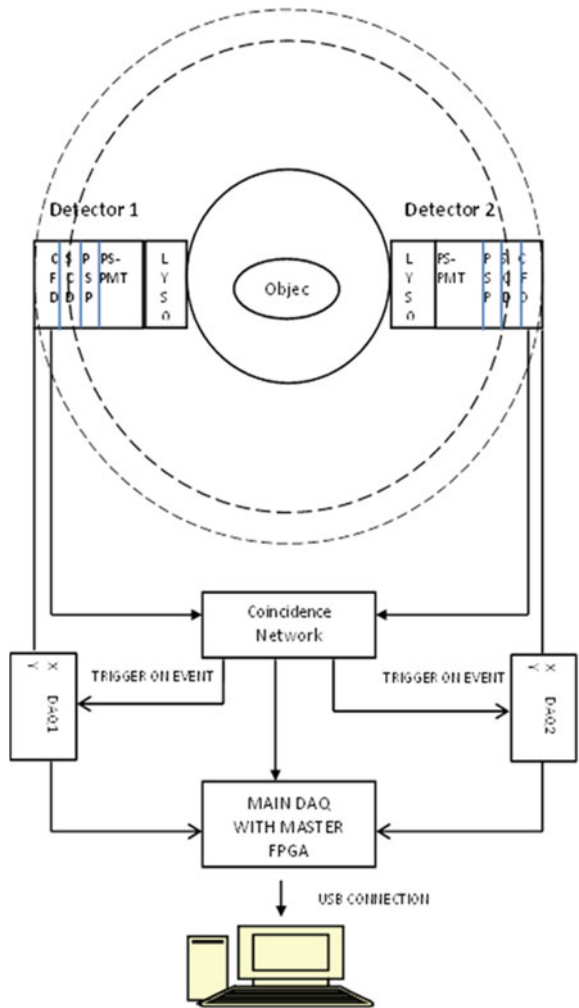
$$\text{OSEM-TV} : x_i^{k-1} = \frac{x_i^{k-1}}{1 + b \sum_{k=1}^m (x_i - x_k)} \sum_{j \in S_n} \frac{a_{ij}y_j}{\left(\sum_{i'=1}^N a_{i'j}x_{i'}^{k-1}\right)} \tag{4}$$

Assuming  $b$  take values over 0.01 and usually less than 2. If  $b$  has a big value this will result in over smoothing while when  $b$  value is small this will result in a better image resolution and an increased noise component.

### 3 Materials and Methods

**System Description.** Figure 1 presents a 2D schema of a rotated dual-head small animal PET. The system detector head consists of a LYSO (Lu0.6Y1.4SiO0.5Ce)

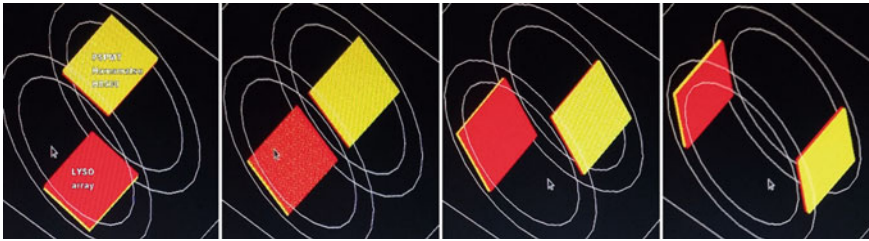
**Fig. 1** A 2D diagram of a prototype dual-head small animal PET system



scintillator  $44.8 \text{ mm} \times 44.8 \text{ mm}$  in size, discretized in  $28 \times 28$  crystals cells. Crystal pixel's dimensions are:  $1.6 \text{ mm} \times 1.6 \text{ mm} \times 1.2 \text{ mm}$ . The scanner has an inner diameter of 160 mm and detectors system is capable of rotating in a gantry between  $0^\circ$  and  $180^\circ$ . Figure 2 shows the camera head rotation over  $90^\circ$ .

LYSO crystals are dominate nowadays together with LSO and other Lutetium variants (LFS-Lutetium Fine Silicate) as a PET scintillator choice. All of them present similar characteristics towards photon detection (Table 1).

Every detector scintillator is followed by a photomultiplier tube (PMT). In the specific PET tomography scheme a Position Sensitive Photomultiplier Tube (PSPMT Hamamatsu H8500) is chosen. After that the signal output of each detector is driven



**Fig. 2** The camera head rotation over 90°

**Table 1** LBS-Lutetium based scintillators characteristics [9]

Characteristics	LYSO	LSO	LFS
Decay time (ns)	53	40	< 33
Light output (APD) (% towards NaI (TI))	85	85	80–85
Peak emission (nm)	420	420	425
Index of refraction	1.81	1.82	1.81
Density (g/cm <sup>3</sup> )	5.37	7.35	7.35
Effective Z	54	65	64
1/μ. mm (511 keV)	20	12.3	11.5
Hygroscopic	No	No	No

further to data processing electronics for online detection of coincidence events during data acquisition procedure [10].

**Phantom Description.** The phantom that was simulated and data from that simulation that is used for the aforementioned study is a Derenzo-like Phantom. The Derenzo-like phantom consists of six different areas of same sized rods in each. The sets of cylinders, are with diameters 4.8, 4, 3.2, 2.4, 1.6, and 1.2 mm. Every set of rods has the same diameter separation between radioactive surfaces. The rods are surrounded with plastic (polyethylene). The Derenzo-type phantom was filled with F<sup>18</sup>. In Fig. 3 a 2D slice, of the Derenzo phantom, is presented.

**System Matrix.** System matrix [11] was derived from an analytical formula as the area of intersection between two lines of response. In Fig. 4 the hall method is presented with the camera heads in a random rotation angle. Only scanner geometrical characteristics were taken into account.

## 4 Results and Discussion

The PET system simulation was performed with SIMSET, acquiring  $18 \times 10^6$  coincidence events. The 3D data was rebinned to a 2D sinogram data set consisting of

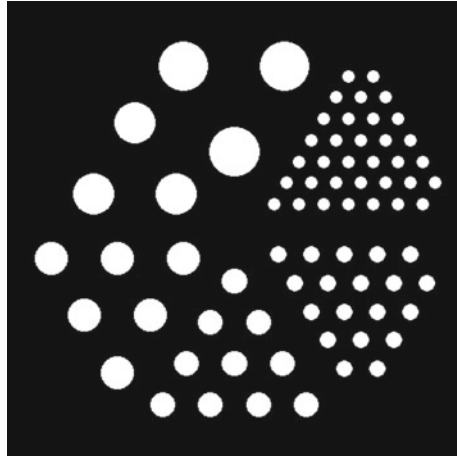


Fig. 3 The Derenzo-like phantom

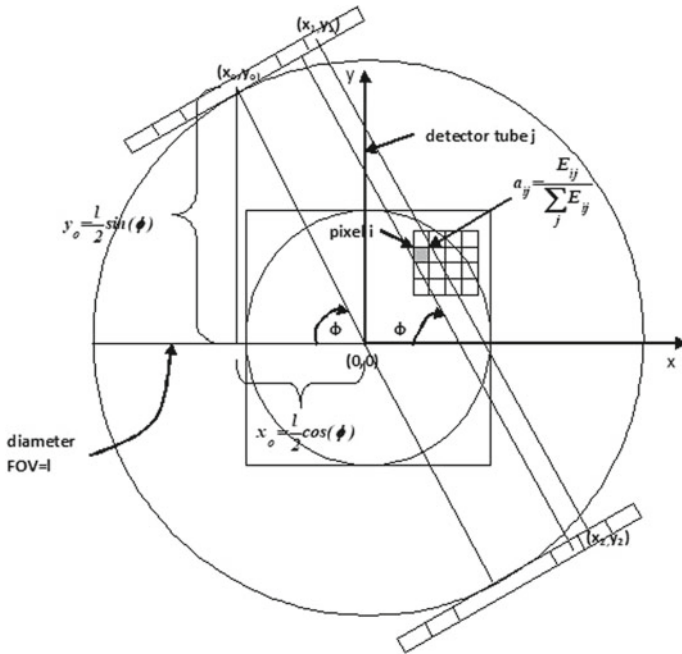


Fig. 4 A schematic view of system matrix element calculation

55pixels × 170 pixels × 28. So, events from 55 LORs were detected and totally 170 angular samples were collected every 1.0647°. In 2D mode 28 parallel and 27 cross LORS were considered to increase system sensitivity. So, coincidences from antidiagonal detector pixels are permitted as well as detector pixels that define LORs with inclination of 0.29°.

The probability matrix was calculated once and stored in disk. It consists of many elements with zero value. The non-zero items were stored. The image size choice was 128 pixels × 128 pixels, so the array A constitutes of 55 × 170 × 128 × 128 items that only 4.33% of them were non-zero in value.

The initial image estimate for OSEM, OSEM-MRP and OSEM-TV is:

$$x_{oi} = \frac{\sum_{j=1}^M y_j}{N^2}, \quad i = 1, 2, \dots, N \tag{5}$$

And the number of subsets is 24 for all reconstruction procedures.

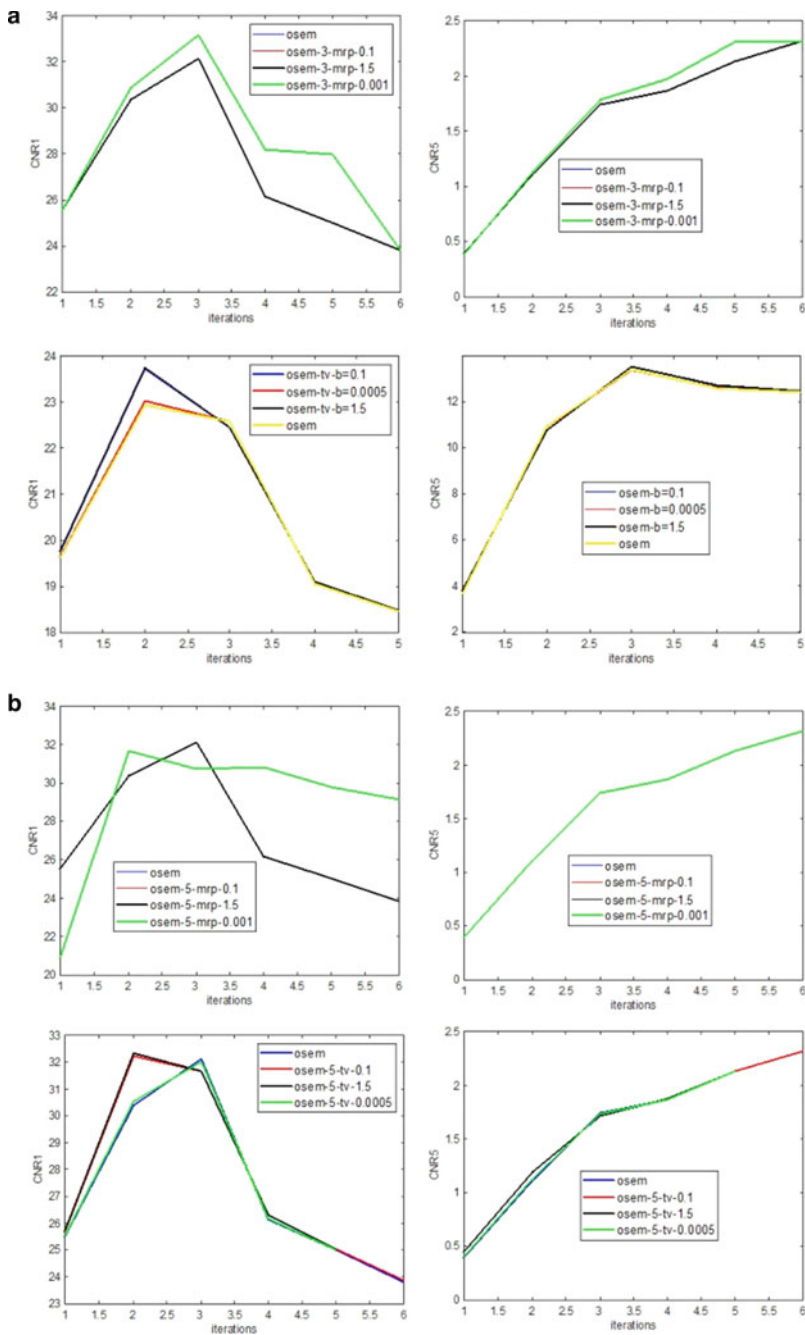
Apart from image profiles, from where the system total resolution feasibility can be derived, local contrast-to-noise ratios (CNR) for rods 4.8, and 1.6 mm in diameter were calculated. CNRs for each cylinder were obtained by assigning squared regions-of-interest (ROIs). The size of ROIs was 4.55, and 2.15 mm, respectively. Inside every rod, in an area of the same objects, a ROI was positioned. Identical-sized ROIs were placed in three different background regions, for the set of same rods. CNR<sub>ROI</sub> was calculated as:

$$\text{CNR}_{\text{ROI}} = \frac{R_{\text{obj}_{\text{ROI}}} - R_{\text{Backg}_{\text{ROI}}}}{\sigma_{\text{Backg}_{\text{ROI}}}} \tag{6}$$

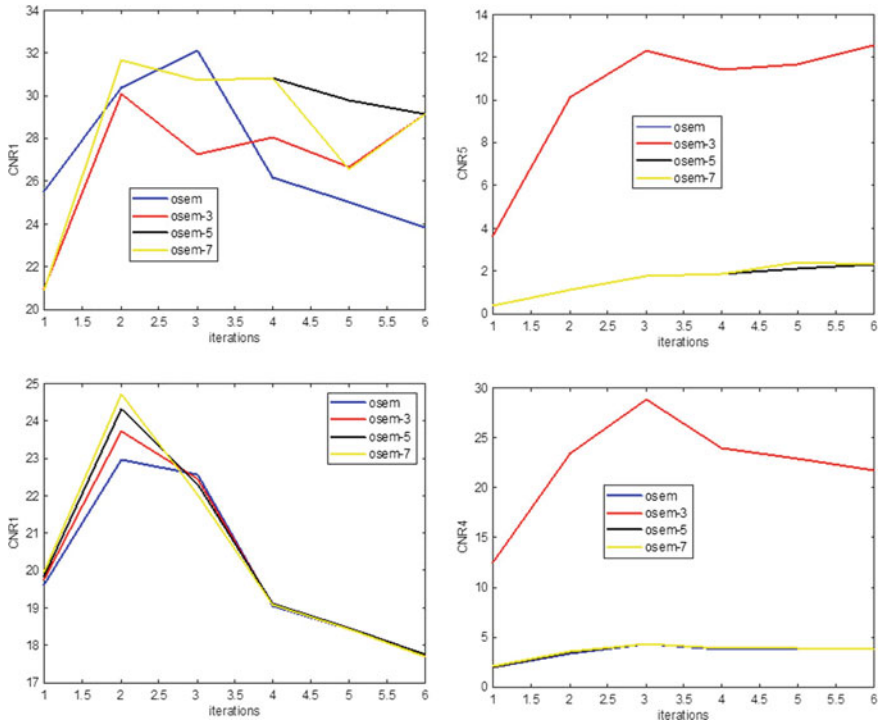
where  $R_{\text{obj}_{\text{ROI}}}$  stands for the average intensity of objects that are reconstructed in all the same diameter cylinder ROIs and  $R_{\text{Backg}_{\text{ROI}}}$  represents the average intensity of the background ROIs in both the case of 4.8 and 1.6 mm diameter rods.  $\sigma_{\text{Backg}_{\text{ROI}}}$  stands for the average background standard deviation in the corresponding ROIs.

In Fig. 5 CNRs are presented for regularization masks 3 × 3 and 5 × 5 both for MRP and TV accompanied with OSEM, for different multiplicative factors b. Figure 5 concerns cylinders with 4.8 and 1.2 diameter. As it can be seen when  $b_{\text{MRP}} = 0.001$  OSEM-MRP produces images with better CNRs for big size objects both for 3 × 3 and 5 × 5 masks. Also, the same multiplicative factor enhances small size objects. The rest values of b for the MRP prior produce no change as far as standard OSEM concerns.  $b_{\text{TV}}$  can be slightly bigger about 0.1. The value 1.5 produces the same results as 0.1, while when  $b = 0.001$  CNRs for OSEM-TV does not differ from OSEM. The value  $b_{\text{TV}}$  presents better contrast to noise ratios for big and small size objects, when mask size is 3 × 3. So for the rest of the study the choice is  $b_{\text{MRP}} = 0.001$  and  $b_{\text{TV}} = 0.1$ .

Figure 6 shows CNRs for OSEM-MRP with  $b = 0.001$  for mask sizes 3 × 3, 5 × 5, and 7 × 7. From the CNR curves it is obvious that a 3 × 3 mask enhances image detail without to increase noise component. The same still for the bottom row



**Fig. 5** CNRs for rods of (left columns) 4.8 mm and (right columns) 1.6 mm for **a** regularization mask  $3 \times 3$  and **b**  $5 \times 5$  mask both for MRP (up rows) and TV (down rows)



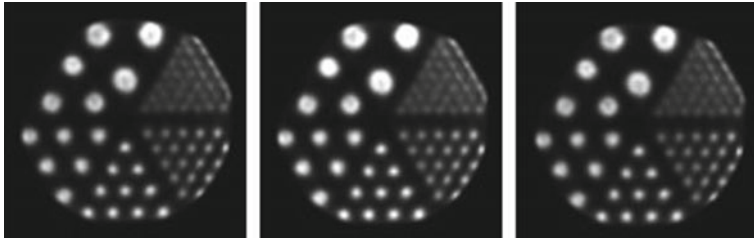
**Fig. 6** Top row) CNRs for OSEM-MRP for different mask sizes, (Bottom row) CNRs for OSEM-TV for different mask sizes for objects of (left column) 4.8 mm and (right column) 1.6 mm diameter

of Fig. 6, where CNRs for OSEM-TV with  $b = 0.1$  for the same three masks are presented.

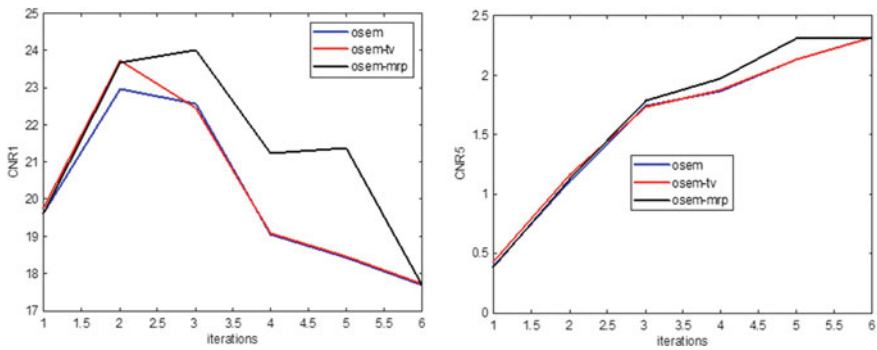
Big size masks  $5 \times 5$  or  $7 \times 7$  enhance image big objects while the small mask of 8 neighbors enhance small objects without increasing noise component. So, a good choice is a small mask. Image detail as small as possible is preferable and helps an early diagnosis.

Figure 7 shows 2D reconstructed images with the standard OSEM, OSEM-MRP (mask  $3 \times 3$ ), OSEM-TV (mask  $3 \times 3$ ) after 3 iterations, while the number of subsets is 24,  $b_{MRP} = 0.001$ ,  $b_{TV} = 0.1$ . Image Reconstruction was performed on an Intel i7 computer by a software written in Visual C++ by the author.

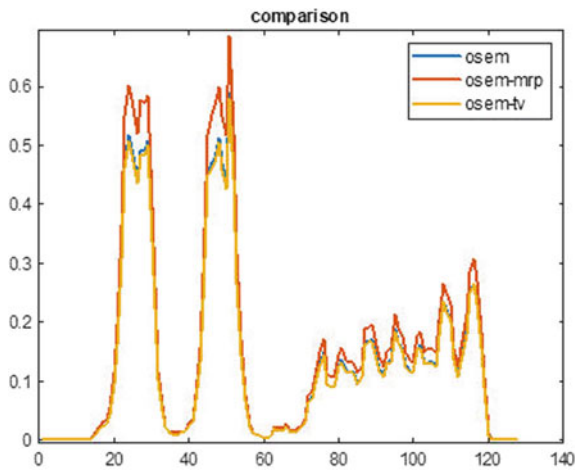
CNRs for OSEM, OSEM-MRP, OSEM-TV for a  $3 \times 3$  mask are presented in the same diagram in Fig. 8. As it can be derived OSEM-MRP is preferable for big and very small objects. As far as image resolution concerns, it can be calculated from the profiles of Fig. 9. So, image resolution calculated for the small objects of 1.2 mm in diameter as the FWHM (Full Width at Half Maximum) of the middle curve is about 2.1 mm for all three reconstruction schemes. This resolution value is similar to clinical and other experimental small animal PET systems, where the resolution value varies from 1 mm to 2.4 mm at the edges of the field of view [3–5, 12, 13].



**Fig. 7** 2D reconstructed images with the standard OSEM, OSEM-MRP (mask  $3 \times 3$ ), OSEM-TV (mask  $3 \times 3$ ) after 3 iterations, while the number of subsets is 24,  $b_{MRP} = 0.001$ ,  $b_{TV} = 0.1$



**Fig. 8** CNRs for OSEM, OSEM-MRP, OSEM-TV for a  $3 \times 3$  mask for objects of 4.8 mm (left) and 1.6 mm (right) diameter



**Fig. 9** Normalized profiles of line 57 for OSEM, OSEM-MRP, OSEM-TV for a  $3 \times 3$  mask



## 5 Conclusions

In this study, Regularization schemes are introduced and evaluated, namely OSEM-MRP and OSEM-TV. These two algorithms are compared towards standard OSEM. Depending on data study regularization mask and multiplicative factor value must to be determined. MRP schemes present a better noise manipulation and keeps image quality in adequate standards.

## References

1. S.R. Cherry, J. A. Sorenson, M. E. Phelps, "Physics in Nuclear Medicine", SAUNDERS-Elsevier, 2003, ch. 15
2. H. M. Hudson and R.S. Larkin, "Accelerated Image Reconstruction using Ordered Subsets of Projection Data", IEEE Trans. Med. Imag. Vol 13, No 4, pp: 601–609, 1994
3. Michał Wyrzykowski, Natalia Siminiak, Maciej Kaźmierczak, Marek Ruchała & Rafał Czepczyński, "Impact of the Q.Clear reconstruction algorithm on the interpretation of PET/CT images in patients with lymphoma", EJNMMI Research (2020) 10:99
4. Syahir Mansor, Elisabeth Pfahler, Dennis Heijel, Martin A. Lodge, Ronald Boellaard, Maqsood Yaqub, "Impact of PET/CT system, reconstruction protocol, data analysis method, and repositioning on PET/CT precision: An experimental evaluation using an oncology and brain phantom", Med. Phys. 44 (12), December 2017
5. Wenyuan Qi, Ting Xia, Xiaofeng Niu, Changguo Ji, Mark Winkler, Evren Asma, Wenli Wang, "A non-local means post-filter with spatially adaptive filtering strength for whole-body PET", 2015 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)
6. P.J. Green, "On Use of the EM Algorithm for Penalized Likelihood Estimation", Journal of Royal Statistical Society Series B, Vol 52, No 3, pp: 443–452.
7. S. Alenius, U Ruotsalainen and J. Astola, "Using Local Median as the Location of the Prior Distribution in Iterative Emission Tomography Image Reconstruction", Proc. IEEE Medical Image Conference 1997.
8. A. B. Jimenez, "Efficient Optimization methods for Regularized Learning: Support Vector machines and Total-Variation Regularization", PhD Thesis, Universidad Autonoma de Madrid, May 2011
9. Qingyang Wei, "Intrinsic Radiation in Lutetium Based PET Detector: Advantages and Disadvantages"
10. Giancarlo Sportelli, "A Modular Data Acquisition System for High Resolution Clinical PET Scanners", PhD thesis 2010
11. R. L. Siddon, "Fast calculation of the exact radiological path for three-dimensional CT array", Medical Physics, vol. 12, pp: 252–255, March 1985
12. Tina Binderup, Henrik H. El-Ali, Valentina Ambrosini, Dorthe Skovgaard, Mette Munk Jensen, Fan Li, Birger Hesse, Jesper Tranekjær Jørgensen and Andreas Kjær, "Molecular Imaging with Small Animal PET/CT", Current Medical Imaging Reviews 7, 234–247, 2011
13. Claudia Kuntner and David Stout, "Quantitative preclinical PET imaging: opportunities and challenges", Frontiers in Physics, REVIEW ARTICLE, published: 28 February 2014

# Early Detection of Parkinson's Disease Dementia Using Dual-Sided Multi-scale Convolutional Neural Networks (DSMS-CNN)



Callum Altham, Huaizhong Zhang, Marcello Trovati, Ella Pereira, Nicola Ray, Simon Keller, Antonella Macerollo, and Hulya Wiesmann

**Abstract** Detecting the potential for Parkinson's Disease Dementia (PDD) as early as possible is crucial to ensure that quality of life can be maintained. However, the full origins of this condition are unknown and analysing potential causes such as the influence of the Cholinergic Basal Forebrain (cBF) can be challenging due to variation in brain tissue as well as low scan resolution. Additionally, the structure and function of the cBF can span both brain hemispheres, and therefore prove difficult to analyse using a singular deep learning method. In this paper, we propose a multi-scale, dual-sided approach to analysis of regions with low surface area such as the cBF. Initially, images are parsed using super-resolution to increase resolution and contrast. Then, a dual sided multi-scale convolutional neural network (DSMS-CNN) model is proposed to classify subjects as either normal cognition or PDD based on both hemispheres of the cBF together. Ablation studies and comparison experiments with state-of-the-art CNN models show that DSMS-CNN can achieve promising and superior performance.

**Keywords** Parkinson's disease dementia · Parkinson's disease · Magnetic resonance imaging · Convolutional neural network

---

C. Altham (✉) · H. Zhang · M. Trovati · E. Pereira  
Edge Hill University, Ormskirk, England, UK  
e-mail: [althamc@edgehill.ac.uk](mailto:althamc@edgehill.ac.uk)

N. Ray  
Manchester Metropolitan University, Manchester, England, UK

S. Keller · A. Macerollo  
University of Liverpool, Liverpool, England, UK

H. Wiesmann  
Liverpool University Hospitals NHS Foundation Trust, Liverpool, England, UK

## 1 Introduction

Parkinson's Disease (PD) is an incurable neurodegenerative disease resulting in unintended, uncontrollable movements such as tremors, as well as issues with coordination [1]. However, those with PD can also develop cognitive complications causing continual, severe neurological decline and loss of cognitive functioning such as Parkinson's Disease Dementia (PDD)[2]. Current research indicates that PDD and traditional Dementia share many clinical, neurological, and morphological features [3] with works seemingly showing that development of PDD can be partially attributed to degradation of tissue structures of the brain, with particular focus on regions that form the cholinergic system of the brain [4]. One main area of interest within this system that has been highlighted for its potential influence on development of cognitive impairments such as PDD is that of the cholinergic Basal Forebrain (cBF) [5]. This area has widespread connections throughout the brain whilst also serving as the primary source of cholinergic innervation to the cerebral cortex [6]. In particular, the cBF has been proven to be important in conditions such as Dementia, PDD and Alzheimer's Disease (AD), with work suggesting that changes to tissue structures and volume of regions within the cBF, namely that of regions Ch1-4p, have potential to identify patients who have mild cognitive impairment (MCI) and are likely to convert to AD or PDD [7], whilst cognitive decline has also been linked to a loss of up to 96% of neurons within region Ch4/Ch4p in PDD patients compared to control groups [5] [8]. Additionally, greater atrophy of these regions can be found in the early stages of PDD compared to controls [9]. This therefore indicates that the analysis of the cBF has potential for use in early detection of PDD.

Magnetic Resonance Imaging (MRI) is commonly used for analysis of brain structures such as the cBF [10] and as a result, analysis of such structures using MRI and deep learning (DL) techniques has given rise to considerable focus [11][12]. However, analysis of areas with such little surface area still face issues due to variation in region presentation, as well as low contrast in images. Some studies aim to address these issues using super resolution techniques [13] to produce an increase in overall image resolution that allows for increased clarity in presentation of smaller regions in MRI [14]. To this end, we implemented super-resolution techniques to improve intensity contrast surrounding the cBF so as to produce more effective analysis. Furthermore, analysing regions that are spread across both hemispheres can be problematic because of an increased separation distance between regions. This therefore inspires us to propose a dual-sided deep learning architecture for analysing data from both brain hemispheres and providing a consolidated prediction as a dual sided, multi-scale convolutional neural network (DSMS-CNN) approach.

There are three main contributions within our work: (1) A multi-scale CNN (MS-CNN) is utilised for the analysis of regions with limited surface area within MR images and extraction of the most important features [15]; (2) A dual classification approach (DSMS-CNN) is proposed to produce a consolidated prediction based on both hemispheres of the cBF at once; (3) Experiments conducted against data

gathered by the Parkinson's Progression Markers Initiative (PPMI) [16] prove that the proposed method is able to provide considerable predictions and has potential to outperform state-of-the-art methods.

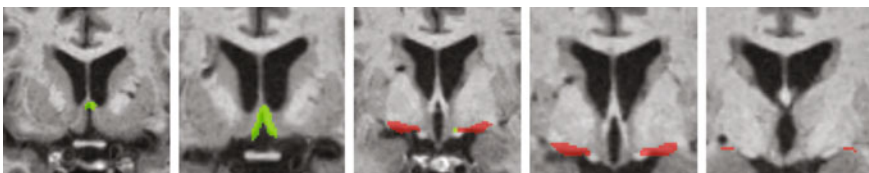
## 2 Proposed Method

### 2.1 Region of Interest

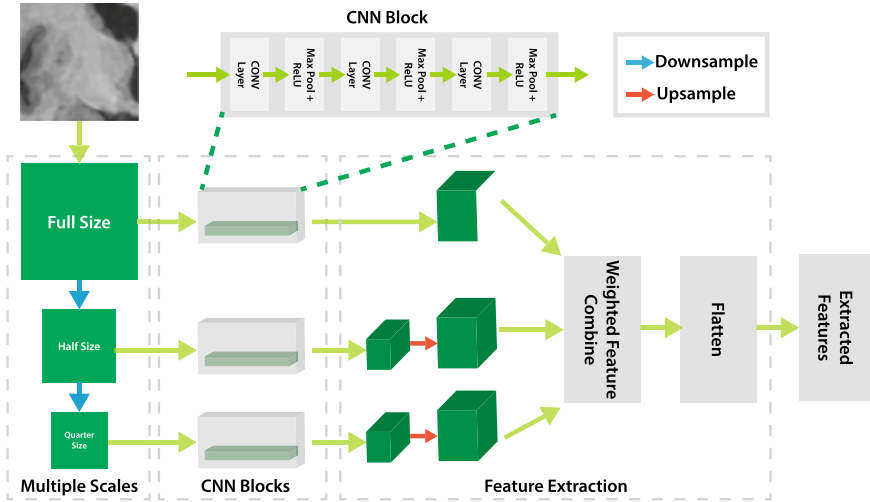
Degeneration of the cBF has been indicated for potential direct links to development of PDD in those with PD [5]. Therefore, this region was chosen for use as a region of interest (RoI) for analysis in this work. Identification of the cBF region came from information gained from the use of a stereotactic map of the cBF [17]. From this map, identification of different subdivisions of the cholinergic system based on Mesulam's nomenclature was possible, with these subdivisions corresponding to the Ch1-4p regions of the cBF [18]. Examples of the RoI before hemisphere separation can be seen in Fig. 1.

### 2.2 Multi-scale Convolutional Neural Network

Computer vision tasks make use of CNNs since they can be designed for learning of abstract and translationally invariant features without many parameters, largely due to influences of successive convolutional layers combined together [19]. However, whilst CNN based methods provide state-of-the-art performance, they are limited in their ability for learning multiscale features based on the number of filters, depth of architecture, and quantity of training data, opening such networks up to possible overfitting [20]. Therefore, to address these issues probably arising in this study, the architecture forming the main building block is the Multi Scale Convolutional Neural Network (MS-CNN) [15] that aims to perform multi-scale classification against the RoIs mentioned above, and has previously demonstrated its efficiency and effectiveness in the analysis of land use data.



**Fig. 1** cBF regions of interest. Green corresponds to regions Ch1-3. Red corresponds to regions Ch4 and Ch4p



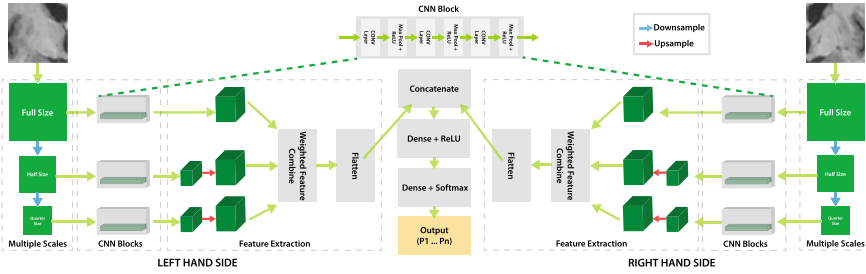
**Fig. 2** Block diagram of the MS-CNN architecture. Image data is represented by green blocks, with any other coloured blocks representing one or more neural layers

In MS-CNN, input images containing the RoI are rescaled to reproduce the RoI at three resolutions, namely full, half and quarter resolutions, meaning that image dimensions provided to the architecture are first halved and then quartered to produce three different inputs for each CNN block within the MS-CNN. An overall diagram of the MS-CNN architecture can be seen in Fig. 2.

### 2.3 Proposed Dual-Sided Architecture

**Classification Problem:** In this study, the main classification problem is that of accurately differentiating between those participants with PD who are at risk of developing PDD within the next 5 years (PD-PDD) and those that are not (PD-NC). This therefore results in the availability of two separate classes, namely PD-NC and PD-PDD.

As mentioned above, cBF regions Ch1-4 occur in both left and right hemispheres of the brain, resulting in differences in feature extraction and model performance that depend on the hemisphere of the brain in question. Therefore, a dual-sided architecture is proposed to analyse each hemisphere independently which allows for proper feature extraction and efficient classification performance. The proposed CNN architecture consists of two independent MS-CNN blocks, as shown in Fig. 3. Each block aims to analyse an individual hemisphere of the cBF and are trained at an image size of  $48 \times 48$ . Features generated from each MS-CNN block are then combined, at which point classification is made against both sides together to produce a singular output of either PD-PDD or PD-NC.



**Fig. 3** Block diagram of the proposed DSMS-CNN architecture. Image data is represented by green blocks, with any other coloured blocks representing one or more neural layers

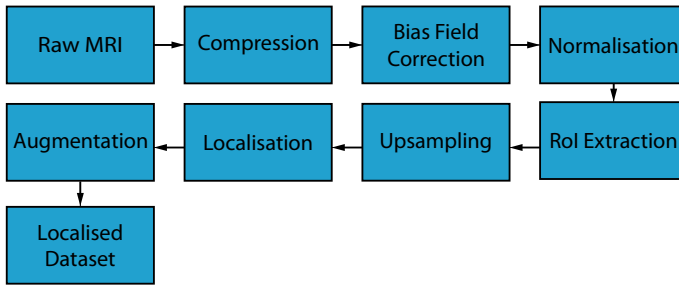
### 3 Experiments and Analysis

#### 3.1 Dataset

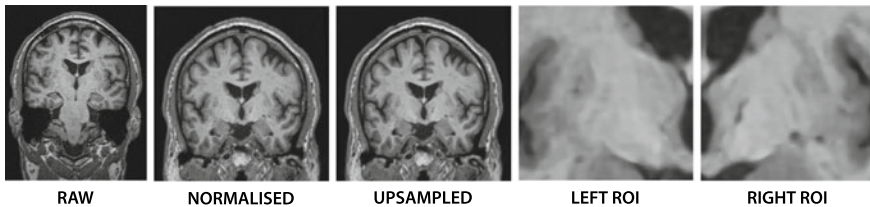
All data used is from the PPMI, which aims to identify markers relevant for tracking of Parkinson’s Disease risk, onset and progression [16]. Subjects selected were those with T1-weighted MRI scans, comprising 386 PD and Healthy Control (HC) subjects. Only those with PD were chosen for use and therefore all HC subject data was discarded due to irrelevance, leaving 288 subjects. These subjects were categorised into either Parkinson’s Disease with “normal” cognition (PD-NC), which is those with a cognitive state that has not degraded to a level considered substantial, or Parkinson’s Disease Dementia (PD-PDD), which is those with severe cognitive decline indicative of dementia. This was done using a set of cognitive and clinical assessments carried out by PPMI, with an overall decision made based on consolidation of 5 years of data per subject after their joining the study. The resulting data suffered from a class imbalance due to 167 PD-NC subjects and 83 PD-PDD subjects. To address this imbalance, the PD-NC group was randomly reduced to the same level as the PD-PDD group, whilst ensuring that group demographics were maintained. This resulted in a final subject pool of 166 subjects equally distributed across both class groups.

#### 3.2 Preprocessing

PPMI data comes from many institutions, so there exists potential for variation and disparity across all data samples. Therefore, all data was processed to ensure it was in a common and standardised format for effective comparison. To perform this pre-processing, the Advanced Normalization Tools (ANTs) [21] and FMRIB Software Library (FSL) [22] were utilised through the Nipype Python programming library [23]. A pipeline of the pre-processing methodology used can be seen in Fig. 4.



**Fig. 4** Processing pipeline



**Fig. 5** Example of pre-processing procedures against MRI scan. Bias field corrected images are identical to raw images and therefore are omitted

All images are first compressed to a usable format for processing. Non-uniformities within image intensities are then removed through the use of N4 bias field correction [24]. After this, all images are aligned to a standardised template space, namely the Montreal Neurological Institute template using FLIRT [25][26]. Based up on the RoI, a set of 9–12 slices purported to contain this RoI were then extracted from each subject data depending on image quality and registration. These slices were considered to be too low of a resolution and were therefore upsampled using super-resolution to be 8x their original size [13]. All images were then ‘localised’ to only contain that of the RoI by cropping each individual slice. This localisation was carried out on both hemispheres to produce 2 images per slice. A lower than preferable number of subjects were gathered from PPMI, so the number of available images was artificially augmented to 7x the number of images. This resulted in a dataset containing around 13,000 images per category. These images were then separated into three separate datasets: left and right hemisphere alone datasets, and a combined dataset containing all available images. Examples of this preprocessing can be seen in Fig. 5.

### 3.3 Parameter Settings

- (1) **MS-CNN** MS-CNN models were trained at a variety of different image sizes to determine the most suitable image size for performance, with an overall size of  $48 \times 48$  pixels performing best. CNN and Feature Extraction blocks are formed of three repeating blocks, within each is a convolutional layer ( $32 \ 3 \times 3$  kernels), a max pooling layer ( $2 \times 2$  window), and ReLU activation.

**Table 1** Number of samples used for training and testing in all models. DSMS-CNN uses both the 'LHS' and 'RHS' datasets independently so are kept separate, whilst all other models are trained on the 'Combined' dataset which includes all images from the LHS and RHS datasets together

Class	LHS	RHS	Combined
PD-NC	6840	6840	13680
PD-PDD	6936	6936	13872

- (2) **Dual-Sided Architecture** To allow for a combined decision making progress, feature maps from both the left and right MS-CNN are combined using a concatenation layer, and the remaining two dense classification layers containing 32 and 2 neurons respectively to form an output.
- (3) **Sample Sizes** From the categorised and extracted participant data introduced in 3.2, the number of samples used for training all models are shown in Table 1.
- (3) **Evaluation Metrics** The prediction performance is evaluated using the metrics of sensitivity, specificity and F1-Score. Sensitivity is to calculate the ability of the model to predict presence of a positive case (A). Specificity is to calculate the ability of the model to predict presence of a negative case (B). F1-Score is the harmonic mean of a models precision and recall score (C).

$$(A) = \frac{TP}{TP + FN} \quad (B) = \frac{TN}{TN + FP} \quad (C) = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (1)$$

where TP, FP, and FN denote true positive, false positive and false negative predictions respectively.

### 3.4 Ablation Study

To examine the benefits gained through the use of the proposed DSMS-CNN, a set of singular MS-CNN architectures are employed to conduct an ablation study on the three derived datasets that are shown in Table 1.

Since all components of the DSMS-CNN are trained at an image size of  $48 \times 48$  pixels, all ablation experiments were conducted on the same size, and the results obtained are shown in Table 2.

It can be observed that whilst the MS-CNN model does perform increasingly well against the left and right hemispheres of the cBF independently, achieving F1-Scores of 93–94%, as well as overall sensitivity and specificity values of around 92–94%, this result is not consistent when faced with a combined dataset of both the left and right hemispheres together, resulting in a significant performance loss of around 3–4% in F1-Score as well as 2–3% in terms of sensitivity and specificity. This evident struggle to classify well between the PD-NC and PD-PDD classes is likely due to



**Table 2** Classification F1-Score and overall model sensitivity and specificity (%) comparison among MS-CNN variations and proposed DSMS-CNN

Metric	MS-CNN <sub>LHS</sub>	MS-CNN <sub>RHS</sub>	MS-CNN <sub>Combo</sub>	DSMS-CNN
F1-Score	93.63	94.32	90.98	<b>97.55</b>
Overall sensitivity	93.08	94.58	91.85	<b>96.25</b>
Overall specificity	92.54	93.49	93.32	<b>98.27</b>

Best results are shown in bold

the fact that when combined, inherent differences between the formation of the left and right hemispheres of the cBF as a result of PDD degeneration produces conflict between the features extracted from the two classes when attempting to perform classification, resulting in an overall under-performance. However, the evident performance benefit of analysing each hemisphere independently indicates the potential for the use of DSMS-CNN, with an overall performance benefit of 6–7% in terms of classification F1 score when compared with the MS-CNN analysing the combined dataset. Additionally, a similar benefit can be seen in overall sensitivity and specificity, with DSMS-CNN able to predict classes correctly at a rate of 5% higher. Both of these results lead us to conclude that a dual-sided approach to classification allows for more accurate and efficient prediction than using a singular MS-CNN faced with all data samples.

### 3.5 Comparison with State-of-the-Art

Further to the ablation study above, we compare our proposed DSMS-CNN model to a number of existing state of the art methods. The results to this are shown in Table 3. For comparison, we employ a number of different CNN architectures to perform classification. The CNN architectures in use are that of the VGG and Inception architectures, more specifically that of the VGG8 [27] and InceptionV3 variations [28], as well as the MS-CNN architecture that forms the building block of the proposed model. All three comparison models are trained against the combined dataset mentioned above comprised of both the left and right hemispheres of the cBF to allow for adequate comparison against the DSMS-CNN. As can be seen in the table, our proposed model shows a clear and evident out-performance compared to other models, with the highest achieving sensitivity and specificity scores of 96 and 98% respectively, a performance benefit of up to 50% more than other models, as well as gains of between 6 and 45% in terms of classification F1-Score. From above, one of the main reasons that this performance difference likely occurs is due to the presence of confusion between the left and right hemispheres of the cBF when PDD is present in a data sample. Additionally, the use of the MS-CNN to form the DSMS-CNN means that a much shallower network is implemented compared to that of the much

**Table 3** Classification F1-Score and overall model sensitivity and specificity (%) comparison among state-of-the-art CNNs and proposed DSMS-CNN

Metric	InceptionV3	VGG8	MS-CNN	DSMS-CNN
F1-Score	52.11	63.84	90.98	<b>97.55</b>
Overall sensitivity	46.55	82.34	91.85	<b>96.25</b>
Overall specificity	57.04	47.72	93.32	<b>98.27</b>

Best results are shown in bold

deeper comparison CNNs, which are seemingly unable to effectively learn from such a limited region, as well as the influence of the use of a multi-scale method of learning, which produces a method that is more suited for the learning of all relevant features within the chosen data area. The benefit of utilising a shallow, multi-scale network is also evident in Table 3 since the performance of the MS-CNN model compared to other models indicates that even in cases of confusion between hemispheres, MS-CNN type architectures are a much more suited alternative to traditional CNNs for analysing this type of region.

Consequently, it is also worthwhile to note that performance of the singular MS-CNN architectures is of a high quality, but suffers from degradation of classification F1-Score when utilising all data, something that is of importance to ensure analysis is placed against the entire cBF region, demonstrating the advantage of using the proposed approach for this particular task. This therefore shows that the proposed approach provides the most stable and highest performing approach of all tested models.

## 4 Conclusions

Analysis and classification of brain regions known for links to PDD have the potential to play a vital role in the ability to pre-emptively identify those PD sufferers at risk of developing PDD. This work aimed to explore potential for novel classification methods to analyse a crucial brain region that has potential links to PDD known as the Cholinergic Basal Forebrain (cBF) and implement a dual-sided model to combine analysis of both hemispheres of the cBF into a singular classification system. Experimentation clearly indicates the potential for the use of the DSMS-CNN in terms of early prediction of PDD development, with findings suggesting that the use of a dual-sided, combined approach allows for effective analysis against all available information without potential confusion and increased complexity. Whilst this work shows potential, it is largely limited due to the quality available within T1 MR imagery, and therefore this limitation needs to be acknowledged. In further work, we would aim to conduct further testing and analysis against differing medical imaging modalities such as that of DTI and PET which show admirable potential in determining effective analysis of brain regions without reliance on image quality [29].

**Acknowledgements** This work was supported by the Health Research Institute at Edge Hill University.

## References

1. M.J. Armstrong, M.S. Okun, *JAMA* **323**, 548 (2020). <https://doi.org/10.1001/JAMA.2019.22360>
2. E.J. Burton, I.G. McKeith, D.J. Burn, E.D. Williams, J.T. O'Brien, *Brain* **127**, 791 (2004). <https://doi.org/10.1093/BRAIN/AWH088>
3. K.A. Jellinger, A.D. Korczyn, *BMC Medicine* **16** (2018). <https://doi.org/10.1186/S12916-018-1016-8>
4. K.A. Jellinger, *International Journal of Neurology and Neurotherapy* **5** (2018). <https://doi.org/10.23937/2378-3001/1410076>
5. J. Gratwicke, J. Kahan, L. Zrinzo, M. Hariz, P. Limousin, T. Foltynie, M. Jahanshahi, *Neuroscience & Biobehavioral Reviews* **37**, 2676 (2013). <https://doi.org/10.1016/J.NEUBIOREV.2013.09.003>
6. E.J. Mufson, S.D. Ginsberg, M.D. Ikonovic, S.T. DeKosky, *Journal of Chemical Neuroanatomy* **26**, 233 (2003). [https://doi.org/10.1016/S0891-0618\(03\)00068-1](https://doi.org/10.1016/S0891-0618(03)00068-1)
7. L. Zaborszky, L. Hoemke, H. Mohlberg, A. Schleicher, K. Amunts, K. Zilles, *NeuroImage* **42**, 1127 (2008). <https://doi.org/10.1016/J.NEUROIMAGE.2008.05.055>
8. K.M. Cullen, G.M. Halliday, *Neurobiology of Aging* **19**, 297 (1998). [https://doi.org/10.1016/S0197-4580\(98\)00066-9](https://doi.org/10.1016/S0197-4580(98)00066-9)
9. J.E. Lee, K.H. Cho, S.K. Song, H.J. Kim, H.S. Lee, Y.H. Sohn, P.H. Lee, *Journal of Neurology, Neurosurgery & Psychiatry* **85**, 7 (2014). <https://doi.org/10.1136/JNNP-2013-305062>
10. A.S. Lundervold, A. Lundervold, *Zeitschrift für Medizinische Physik* **29**, 102 (2019). <https://doi.org/10.1016/J.ZEMEDI.2018.11.002>
11. A.M. Dawud, K. Yurtkan, H. Oztoprak, *Computational Intelligence and Neuroscience* **2019** (2019). <https://doi.org/10.1155/2019/4629859>
12. J.S. Paul, A.J. Plassard, B.A. Landman, D. Fabbri, (2017), p. 1013710. <https://doi.org/10.1117/12.2254195>
13. C. Zhao, M. Shao, A. Carass, H. Li, B.E. Dewey, L.M. Ellingsen, J. Woo, M.A. Guttman, A.M. Blitz, M. Stone, P.A. Calabresi, H. Halperin, J.L. Prince, *Magnetic Resonance Imaging* **64**, 132 (2019). <https://doi.org/10.1016/J.MRI.2019.05.038>
14. Y. Li, B. Sixou, F. Peyrin, *IRBM* **42**, 120 (2021). <https://doi.org/10.1016/J.IRBM.2020.08.004>
15. H. Zhang, C. Altham, M. Trovati, C. Zhang, I. Rolland, L. Lawal, D. Wegbu, N. Ajiienka, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **15**, 7631 (2022). <https://doi.org/10.1109/JSTARS.2022.3203234>
16. K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kiebertz, E. Flagg, S. Chowdhury, W. Poewe, B. Mollenhauer, T. Sherer, M. Frasier, C. Meunier, A. Rudolph, C. Casaceli, J. Seibyl, S. Mendick, N. Schuff, Y. Zhang, A. Toga, K. Crawford, A. Ansbach, P. de Blasio, M. Piovella, J. Trojanowski, L. Shaw, A. Singleton, K. Hawkins, J. Eberling, D. Russell, L. Leary, S. Factor, B. Sommerfeld, P. Hogarth, E. Pighetti, K. Williams, D. Standaert, S. Guthrie, R. Hauser, H. Delgado, J. Jankovic, C. Hunter, M. Stern, B. Tran, J. Leverenz, M. Baca, S. Frank, C.A. Thomas, I. Richard, C. Deeley, L. Rees, F. Sprenger, E. Lang, H. Shill, S. Obradov, H. Fernandez, A. Winters, D. Berg, K. Gauss, D. Galasko, D. Fontaine, Z. Mari, M. Gerstenhaber, D. Brooks, S. Malloy, P. Barone, K. Longo, T. Comery, B. Ravina, I. Grachev, K. Gallagher, M. Collins, K.L. Widnell, S. Ostrowizki, P. Fontoura, F.H. La-Roche, T. Ho, J. Luthman, M. van der Brug, A.D. Reith, P. Taylor, *Progress in Neurobiology* **95**, 629 (2011). <https://doi.org/10.1016/J.PNEUROBIO.2011.09.005>
17. I. Kilimann, M. Grothe, H. Heinsen, E.J.L. Alho, L. Grinberg, E. Amaro, G.A.B.D. Santos, R.E.D. Silva, A.J. Mitchell, G.B. Frisoni, A.L. Bokde, A. Fellgiebel, M. Filippi, H. Hampel,

- S. Klöppel, S.J. Teipel, *Journal of Alzheimer's Disease* **40**, 687 (2014). <https://doi.org/10.3233/JAD-132345>
18. M.M. Mesulam, E.J. Mufson, B.H. Wainer, A.I. Levey, *Neuroscience* **10**, 1185 (1983). [https://doi.org/10.1016/0306-4522\(83\)90108-2](https://doi.org/10.1016/0306-4522(83)90108-2)
  19. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, (2015), pp. 1–9
  20. Y. Chen, H. Jiang, C. Li, X. Jia, P. Ghamisi, *IEEE Transactions on Geoscience and Remote Sensing* **54**, 6232 (2016). <https://doi.org/10.1109/TGRS.2016.2584107>
  21. B.B. Avants, N.J. Tustison, M. Stauffer, G. Song, B. Wu, J.C. Gee, *Frontiers in Neuroinformatics* **8** (2014). 10.3389/FNINF.2014.00044/PDF. [/pmc/articles/PMC4009425/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4009425/) [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4009425/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4009425/?report=abstract)
  22. M. Jenkinson, C.F. Beckmann, T.E. Behrens, M.W. Woolrich, S.M. Smith, *NeuroImage* **62**, 782 (2012). 10.1016/J.NEUROIMAGE.2011.09.015. <https://linkinghub.elsevier.com/retrieve/pii/1053811911010603>
  23. K. Gorgolewski, C.D. Burns, C. Madison, D. Clark, Y.O. Halchenko, M.L. Waskom, S.S. Ghosh, *Frontiers in Neuroinformatics* **5**, 13 (2011). <https://doi.org/10.3389/FNINF.2011.00013/ABSTRACT>
  24. N.J. Tustison, B.B. Avants, P.A. Cook, Y. Zheng, A. Egan, P.A. Yushkevich, J.C. Gee, *IEEE Transactions on Medical Imaging* **29**, 1310 (2010). <https://doi.org/10.1109/TMI.2010.2046908>
  25. M. Jenkinson, S. Smith, *Medical Image Analysis* **5**, 143 (2001). 10.1016/S1361-8415(01)00036-6. <https://pubmed.ncbi.nlm.nih.gov/11516708/>
  26. M. Jenkinson, P. Bannister, M. Brady, S. Smith, *NeuroImage* **17**, 825 (2002). 10.1016/S1053-8119(02)91132-8. <https://pubmed.ncbi.nlm.nih.gov/12377157/>
  27. K. Simonyan, A. Zisserman, 3rd International Conference on Learning Representations (2015). 10.48550/arxiv.1409.1556. <https://arxiv.org/abs/1409.1556v6>
  28. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2016-December**, 2818 (2016). <https://doi.org/10.1109/CVPR.2016.308>
  29. A. Khvostikov, K. Aderghal, J. Benois-Pineau, A. Krylov, G. Catheline, arXiv preprint (2018). 10.48550/arxiv.1801.05968. <https://arxiv.org/abs/1801.05968v1>

# A Change Detection with Machine Learning Approach for Medical Image Analysis



Mauro Mazzei 

**Abstract** The research activity focuses on evaluating data from medical images by applying clustering techniques on extracted components for an ex-ante/ex-post evaluation, commonly identified as “change detection”, with respect to evolution times and/or comparison with other analyzed subjects. The methodological approach examines an unsupervised automatic method through the implementation of an algorithm for the extraction of meaningful features related to the properties of the medical image data analyzed. It then goes on to normalize the data contained in the matrix to evaluate through multivariate analysis the notion that similar objects produce similar responses without knowing their entity, type, and class descriptions, which are inferred by making observations on the clusters. The main specificity of this algorithm is that classes are identified from compact, well-distinguishable clusters without knowing the extent of their nature; the entire feature space is divided into classes using proximity or similarity criteria. After finishing the process of class identification, properties will be associated with them in relation to the known descriptions. At the end of the procedure, factorial analysis using the principal component method is applied. The clusters extracted from the data are described by their properties, which make it possible to identify, on each of the new factorial axes, homogeneous classes of clusters characterized predominantly by the only variables that have a high correlation value between variable and factor. The automatic identification of classes or “phenomena” that exhibit very different mean and/or variance allows easy reading of the results for domain experts.

**Keywords** Artificial intelligence · Machine vision · Image processing · Medical diagnosis

---

M. Mazzei (✉)

Istituto di Analisi dei Sistemi ed Informatica “Antonio Ruberti”, Via di Taurini, 19—00185 Rome, Italy

e-mail: [mauro.mazzei@iasi.cnr.it](mailto:mauro.mazzei@iasi.cnr.it)

# 1 Introduction

## 1.1 Basics of Change Detection

The automatic diagnosis of pathologies from biomedical images such as X-rays or MR images (i.e., obtained by Magnetic Resonance) can also be seen as a sub problem of the more general problem of Change Detection (CD), i.e.: the diagnosis of a “change” occurring between two different instances of the same data set (typically two measurements, at different times but with the same measuring instrument, on the same physical object). In the case, here of primary interest, of images, the data set consists, in the most basic case, of brightness values associated with each point in a region of plane (2D image) or space (3D image). In the biomedical field, the use of 4D images is not uncommon, meaning a family of 3D images each associated with an instant of time, taken at some predefined interval. The CD problem consists in the construction of an algorithm that, given two images of the same scene at different instants, produces a binary output (0–1) corresponding to a ‘significant’ difference between the two images (appearance or disappearance of a given object in the image) and that has certain robustness properties with respect to ‘non-significant’ changes (referred to in the English-speaking literature as artifacts, false flags, outliers, etc.). In the case of biomedical images, the binary output, in addition to signaling onset/non-onset of a given pathology (e.g., a tumor), can also signal (this is the case with 4D images) more complex changes such as the progress/regression of the pathology. In the first case, the Change Detector (we will use the acronym CD for this last noun as well) operates on a pair of images of the same patient at different times, one of which is with certainty relative to the patient under disease-free conditions. In the second case, CD operates on two image sequences at different times, one sequence relating to the patient, and the other relating to a typical case of disease course.

CD methodologies can be grouped into two categories: deterministic methods and statistical (or ‘probabilistic’) methods. Both refer to a definition of an image as a set of random variables linked to regions of the plane, in the case of 2D images.

In general, when comparing two images,  $I_1$  and  $I_2$ , (in this report we will not discuss the above-mentioned case of comparing two sequences of images) in which the same object is depicted, and the CD is to detect the possible appearance of some detail in  $I_2$ , under the (ideal) assumption that everything except the detail (i.e., the background) remains unchanged between the two images, it is convenient to operate on the difference image:  $ID = |I_2 - I_1|$ , and indeed CD, under such ideal conditions, easily accomplishes its task, calculating ID as the sum  $S$  of the differences between all the corresponding pixels (or voxels) of the two images.

Most of the methods of which we will give a bibliographical account follow the basic methodology described above and differ from each other solely in the different approach with which they attempt to handle the non-ideality inevitably present in the real cases, namely the fact that:

1. the background may change between  $I_1$  and  $I_2$ ,

2. the object may be differently illuminated in the two images
3. the object may not be perfectly aligned in the two images, i.e., that translation, rotation, enlargement, and/or reduction may be required to perfectly match the object in I2 with the object in I1, and finally
4. the images may be corrupted by noise, i.e., random and mutually independent variations on individual pixels that make the comparison, i.e., the difference image, nonsignificant.

## 1.2 *Deterministic Methods*

In the presence of noise, the pixels in the difference ID image will reproduce all the random alterations present in I1 and I2. The only (deterministic) way to correct for these alterations is to set a threshold  $r > 0$ , and replace the condition  $S(RD) > 0$  with  $S(RD) > r$ . This simple method can give satisfactory results if the random alterations in brightness rarely exceed the threshold value  $r$  and that, in addition, the detail in the object to be detected has brightness values predominantly greater than the threshold itself. For more critical situations deterministic methods must be replaced by the more sophisticated probabilistic methods. These will be accounted for later, but it should be emphasized already now that probabilistic methods often replicate deterministic ones, in the sense that, without changing them in substance they make them more 'robust' by substituting random variables with appropriate distributions for deterministic quantities.

Another very common example of deterministic manipulation is in relation to the problem (3) described above of 'alignment': instead of the ID difference image, the I3 image obtained as a minimize of a certain function, based on the difference, is considered, however, involving the alignment of the two images through variations of rotation, translation, and magnification/reduction operators, easily obtained with software on the market.

It is worth noting that such preprocessing is also often necessary for probabilistic methods. Let us also emphasize that the normalization now described can, in the simplest cases, also serve as a solution to problem (2) listed above, for example, when the different illumination of the object is nevertheless always uniform in each image. If the illumination difference is not uniform the I3 image will still show spurious (no significant) elements due to this no uniformity. As for the background, it is necessary for it to be unchanged and uniformly illuminated to avoid the proliferation of spurious elements. This condition is verified in the cases of interest here of biomedical images obtained by MR or radiography, in which the subject is always imaged in a laboratory environment.

### 1.3 Probabilistic Methods

There In probabilistic methods, the value of the difference image  $ID = I_2 - I_1$  (simple difference, not in modulus) at the generic pixel  $(i, j)$  is generally modeled as a random variable with Gaussian distribution, extracted from images taken from a sample of healthy patients, that is, under the assumption of no change. To identify the region (or regions) containing the ‘pathological detail,’ one proceeds in a manner like what has already been seen in the deterministic case: one scans the difference between the two patient images for regions where the significance test based on the Gaussian distribution identified above.

An a priori probabilistic model for the image can be obtained by considering not only  $H_0$  but also the complementary hypothesis  $H_1$  (presence of change, either in a pixel or in a sub region).

Examples and a thorough description of MOG methods can be found in [1–3]. For other methods, not described in this report, see [4] (Kernel Density Estimation, KDE) [5, 6] (sparse modeling) [7, 8] (compressed sensing), [9] (Sparse dictionaries).

## 2 Change Detection Applied to Biomedical Imaging

This paper aims to provide the results obtained on “CD” techniques used to analyze and classify biomedical images for early diagnosis of neurodegenerative diseases. “CD” techniques refer to the process of identifying changes in a cluster or phenomenon, which occur in a particular time interval. It is of paramount importance in “CD” analyses performed in this context, to assess that the change in image clusters observed by analysis of data detected by positron emission tomography or PET corresponds to a change in radiometric spectral response, and that this spectral change is significantly more relevant than changes due to other factors, such as: boundary conditions at the time of acquisitions, non-congruent placements and/or projections, differences in the acquisition conditions of the detected data.

Comparison, Post classification, Multi-Temporal Data Classification, Principal Component Analysis, Temporal Differences, Change Vector Analysis, are some of the comparison techniques already experimented in the spatial domain. The research activity focused the results on PET image data analysis by providing clusters that summarize the extracted components for ex ante/ex post evaluation with respect to evolution times and/or comparison with other analyzed subjects. This method implements an unsupervised automatic machine learning algorithm capable of selecting the most significant features of the analyzed images and extracting meaningful numerical attributes related to the properties of the selected features. The normalization phase of the data contained in the matrix, subsequently involves a multivariate analysis of the data by exploiting the concept that similar objects produce similar responses without knowing their magnitude, which is why the algorithm is focused on unsupervised machine learning techniques. The peculiarity of the method lies in identifying classes



from compact clusters that are well distinguishable from each other without knowing to the extent of their nature. All content describing the features of the elements is separated into classes using proximity or similarity criteria. After the class identification phase is finished, the properties are associated in relation to the descriptions made through the numerical indices applied. Next, factorial analysis using the principal component method is applied. The clusters extracted from the data describe their properties, this technique allows their characteristics to be identified on each of the new factorial axes through only those variables that have a high correlation value between variable and factor. This automatically analyzes the classes or “phenomena” with the highest or lowest variance.

Change detection (CD) has always been a subject of study in various fields, such as image surveillance, remote sensing, medical imaging, etc.

The challenge of change detection in medical imaging favors a very objective assessment of the stages of change over time verified through medical imaging diagnoses of the studied pathology. Some of the main challenges lie in the solutions of eliminating all the elements that contribute to noise in the data, and to the change in the patient’s position during the image acquisition phase.

In this area, existing change detection methods are reviewed based on the problems to be addressed and mathematical limitations. Next, the solutions adopted for subspace optimization to approximate the image background more efficiently are presented.

These techniques are based on the main components of structure analysis, with the goal of capturing the similarity of values between two acquisitions in the context of change detection. We discuss theoretically and numerically the choices made and used in subspace approximation.

The mathematical approaches developed next consist of:

- A new mathematical model for change detection, defining it as a clustering problem having a set of data (observation, points, vectors.), one must try to group such data into subsets, a subspace and a similarity measure.
- Development and implementation of numerical pipelines to calculate clinical changes by designing mathematical algorithms.
- Optimization of algorithms by introducing a global co-registration.
- Introduce two new subspace structure learning models that are robust to outliers and noise, reduce the dataset size, and compute actively and efficiently.

The co-registration phase was defined as a minimization problem, all elements that are less than a certain threshold are eliminated as they are influential in the data evaluation phase.

Based on the applied mathematical models, numerical algorithms have been developed that allow class identification so that clinically unrelated changes are automatically excluded and true changes in structure that may be of clinical importance are identified.

The approaches are data-driven and use machine learning or Machine Learning knowledge.

The approach is based on quantitatively and objectively analyzing the performance of these algorithms using synthetic, real-world data. The work presents potential for use in computer-aided diagnosis.

### 3 Related Work

Change detection (CD) algorithms are utilized to identify regions of change in multiple images of the same scene taken at different times.

These techniques have undergone continuous development over the years and, today a variety of algorithms, methods and automated systems are used mostly these algorithms are based with a deterministic approach and few utilize a probabilistic approach.

In clinical practice it is of great importance, the detection of changes in medical images taken at different times.

For all imaging modalities such as MRI, computed tomography, etc., here we mainly focus on MRI.

Imaging datasets may have multiple sequences, each also consisting of many images, these images must be compared with the immediately preceding study or multiple previous studies obviously always referring to the patient examined.

The main problem of change detection algorithms in serial MRI images is to be able to detect changes that detect the pathology being studied, discarding data that are not needed, such as those induced by noise and that are not meaningful to the diagnosis. In addition, misalignment of data can cause much annoyance and result in errors in assessments.

With respect to the diversity of approaches used, a change detection algorithm usually consists of many common preprocessing steps especially in the techniques of suppressing or filtering non-significant changes that are usually detected in the boundary conditions of the evaluated pathology. These preprocessing techniques make it possible to evaluate the set of pixels that are significantly different from previous images and are related to the disease.

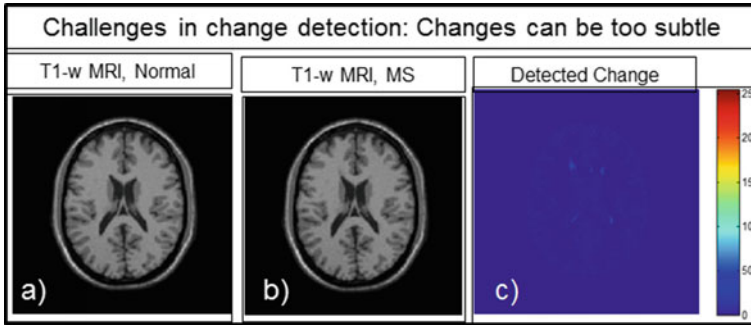
In medical diagnosis and treatment, serial MRI examinations are often performed on patients in the field of neurodegenerative diseases, in monitoring on patients with diseases such as cancer, multiple sclerosis, etc.

Radiologists routinely detect small changes in images of the same anatomical location that may be clinically significant, as shown in Fig. 1.

Detecting small changes in extension or character, using a side-by-side presentation mode, can be very tedious and cause gross errors.

Radiologists try to correct with the tools made available to visually verify errors introduced by patient repositioning and use their professional knowledge to identify and reject certain detections.

Challenges radiologists face during visual comparison include:



**Fig. 1** **a** T1-weighted MRI of a normal brain. **b** T1-weighted MRI of a brain with MS lesions, **c** image containing only disease-related changes. (Images from <http://mouldy.bic.mni.mcgill.ca/brainweb>)

- Use of scanner software and hardware-related parameters (pulse sequences, acquisition parameters, gradient, RF inhomogeneity, registration).
- Separation of acquisition changes from disease-related changes.
- Overload of data and information.
- Inability to detect changes in framed objects or scenes and inability to make comparisons between two scenes.
- Satisfaction in the study.

The change occurs in an unexpected location.

Change occurs in a part of a complicated lesion.

The side-by-side presentation is not suitable for proper interpretation.

The use of appropriately designed computerized automated systems can contribute to an improvement for clinical interpretation.

Patriarche, J.W., Erickson, B.J. stated that automated, computerized change detection can be a valuable tool for radiologists.

MRI examinations of brain tumors have been studied with the aim of significantly reducing the human error that can be caused and due to the enormous amount of data that were studied, and automated techniques were used to improve the results. The main purpose of these studies was to reduce human error by minimizing the enormous amount of data that radiologists must process to reach a conclusion.

The same authors found that the implementation of a scientifically useful tool is clinically feasible when it is dynamically integrated into clinical work, concluding that automatic change detection can improve efficiency, accuracy, and agreed-upon sharing of interpretation.

A computerized system that automatically reduces the amount of data and directs radiologists' attention to clinically relevant areas would therefore be very useful.

The automated change detection system created was a great improvement over the previous automated system, however, the process is inherently lengthy, and the task of tissue classification remains very difficult.

Radke, R.J., Andra S., Al-Kofahi O., Roysam B. reviewed many change detection algorithms and classified them into two groups, statistical analysis and basic modeling techniques.

Bosc, M., Heitz, F., Armspach, J.P., Namer, I., Gounot, D., Rumbach, L. presented an automatic change detection system for serial MRI with applications in multiple sclerosis follow-up.

One of the statistical analysis methods used is based on the use of multimodal information for change detection, generalized likelihood ratio test, and normalization of the nonlinear joint histogram: The performance of the algorithm is low when the noise is nonstationary [9–12].

The work of Patriarche, J.W., Erickson, B.J., in “A Review of the Automated Detection of Change in Serial Imaging Studies of the Brain” is perhaps among the best recognized statistical analysis methods in the medical field.

Researchers have implemented an integrated system for change detection in multispectral serial MRI examinations based on post-classification of image pixels in the space of multispectral MRI intensity functions.

Their rationale for the use of multispectral space was based on the observation that abnormal tissue can “look like” a tissue transitioning from one normal tissue to another in feature space and the hypothesis that changes tend to occur along lines connecting pairs of centroid clusters in feature space.

The detected changes were presented in the form of a color-coded change map superimposed on the anatomical images.

The system also formats the output as a quantitative summary. Preliminary clinical studies tend to show that the system can visually identify subtle disease-related changes. However, the task of classifying tissues is itself very difficult; moreover, the entire process of calculating transitional tissue types and fractional membership for each pixel is inherently time-consuming [13, 14].

Another variation detection method for MR range images is proposed by Seo, H.J., Milanfar, P., in “A Non-Parametric Approach to Automatic CD in MRI Images of the Brain using a non-parametric general statistical method based on local processing kernel.

The calculation of the test statistics was derived from cosine similarity. Their work does not address co-registration and is also limited to one imaging modality.

Among the background modeling methods, background subtraction has mainly been used which is derived from classic video surveillance applications where images subtracted from the background are recovered using compression detection (CS).

This method works when the major changes occupy a small part of the test image and therefore the modified image is dispersed in the spatial domain.

Assuming both background and foreground meet the criteria poorly, they solve the problem using minimization with the total variation algorithm.

Other groundbreaking work proposed the use of the principal component in search of a method to detect changes in the foreground. The work is based on the matrix decomposition of low-level and sparse images.

Robust principal component analysis (PCA) has applications in many other areas, such as facial recognition, etc.

This method is applied to a series of video frames and can also be used for a series of MR images [15, 16].

Others use robust dictionary learning to solve the background subtraction problem. Their approach appears to produce a better dictionary than the more traditional approach using the K-SVD algorithm.

However, the same assumptions for scarcity must hold here as well.

Any application of CS to background subtraction models involves the use of various minimization algorithms. Many MRI reconstruction techniques use compression sensing methods.

Incidentally, compression work is well known for direct application to MR images of the brain. It uses a well-known fact that MR images are poor on some domains such as wavelets, finite differences, etc. An underprinted MR image is recovered using minimization, which allows for faster MR imaging.

Aharon, M., Elad, M., Bruckstein A., K-SVD applied dictionary learning techniques to solve the reconstruction problem. They proposed a patch-based adaptive scarifying learning dictionary that is obtained using k-space data and is used to remove aliasing and noise. The dictionary is created using the K-SVD algorithm and updated for each image block.

One of the major challenges in change detection algorithms for medical imaging is to detect disease-related changes by rejecting changes caused by noise and acquisitions of changes such as skew and intensity in homogeneity.

Despite the diversity of approaches, an image-to-image change detection algorithm consists of many common preprocessing steps to filter out insignificant changes before making change detection decisions. The main algorithm is then used to determine the set of pixels that are significantly different from the initial reference image and are related to the disease.

The preprocessing steps complicate the consistency of the algorithm as a whole and increase the calculation time, in these cases it can distort the clinical relevance of the information in the images.

The work of Nguyen, L.H., Tran, T.D. has addressed mismatches in the change tracking problem using a set of optimization problems. Their method only works well for certain types of images, such as synthetic aperture radar (SAR) images which are much rarer than most medical images.

Turk, M., Pentland, A., Needell, D., JA Tropp, JA, CoSaMP Eigen faces for recognition and pattern recognition of principal components of the face distribution or eigenvectors of the covariance matrix of the 'set of face images, where each image is treated as a vector. The eigenvectors are called eigen faces, and each of them represents a different amount of variation between the images. Each test image is compared to many trainings reference images from a database.

These challenges motivated the design of three algorithms that automatically tolerate acquisition-related noise and changes, and capture subtle, clinically important changes in the differences between two or more medical images.

## 4 Case Study

PET data (Positron Emission Tomography) are very useful for obtaining functional images of the patient's organs. The resulting data does not concern only the shape of organs and structures of the body, i.e., the anatomy, but also their metabolism and "functioning". PET data exploit the decay of special radiopharmaceuticals composed of positron-emitting radionuclides bound to specific molecules. The radiopharmaceutical is administered to the patient intravenously and is distributed throughout the body.

Thanks to the so-called "ferry" molecule, the radiopharmaceutical accumulates in a highly selective way in certain parts of the body, after a pre-established time the decomposing radiopharmaceutical emits radiation called positrons.

A positron is a particle like an electron but with opposite electric charge, when a positron meets electron photons are produced and the PET scanner is able to record these photons and transform them into images [17].

The coding standard for images is the DICOM (Digital Imaging and Communications in Medicine) format which defines the criteria for communicating, displaying, archiving, and printing biomedical information. The DICOM standard is public and accessible to all through special data reading software. Its dissemination proves to be extremely beneficial because it provides a basis for the exchange of information between equipment from different manufacturers, especially in the medical environment.

Radiological data represented as images that are stored according to the DICOM standard in the form of files are called DICOM images. The data in the DICOM format does not undergo compression as in the classic JPEG and GIF formats. The DICOM standard applied to file encoding is a method for encapsulating data and defining how it should be encoded or interpreted.

Biomedical images stored in DICOM format take on diagnostic value and have legal value.

The DICOM (Digital Imaging and Communication in Medicine) standard defines how biomedical images and related metadata are stored and transferred between various devices such as scanners, workstations, and servers. The DICOM format can be used to archive data obtained by various methods of biomedical imaging to obtain a complete description of a diagnostic study. Includes radiological imaging methods such as computed tomography (CT), PET, magnetic resonance (MRI), X-ray and ultrasound images.

For this work, PET images were processed by the protocol of the Alzheimer's Disease Neuroimaging Initiative (ADNI), a worldwide project that provides reliable clinical data for research on the pathological principles, prevention, and treatment of Alzheimer's disease [18].

The main objectives of ADNI are the development of improved methods and uniform standards for the acquisition of data, magnetic resonance imaging (MRI), positron emission tomography (PET), on patients affected by neurodegenerative diseases.

The project plan is to develop an accessible data repository describing longitudinal changes in brain structure and metabolism during the acquisition of parallel clinical, cognitive, and biochemical data.

Furthermore, we want to develop methods that maximize the power to determine treatment effects in clinical trials and to test a variety of hypotheses based on clinical data and biomarkers.

Patient control data is sampled over several months, in this example we sampled data at 12 and 24 months. The following is shown in Figs. 2 and 3.

First processing to discriminate significant elements of a PET image cluster and check a posteriori which one is presumed to be closest to the association of a diagnosis interpretation. The test involves evaluating the results on a series of PET images made at different times of the same subject, keeping the boundary conditions unchanged, see Fig. 4 for reference image.

The method of analysis of the PET image that I propose involves the definition of numerical algorithms that examine all the sections identified for a correct evaluation and interpretation. The objective to be achieved is to recognize some of the information collected in the form of a matrix to obtain dynamic features in order to

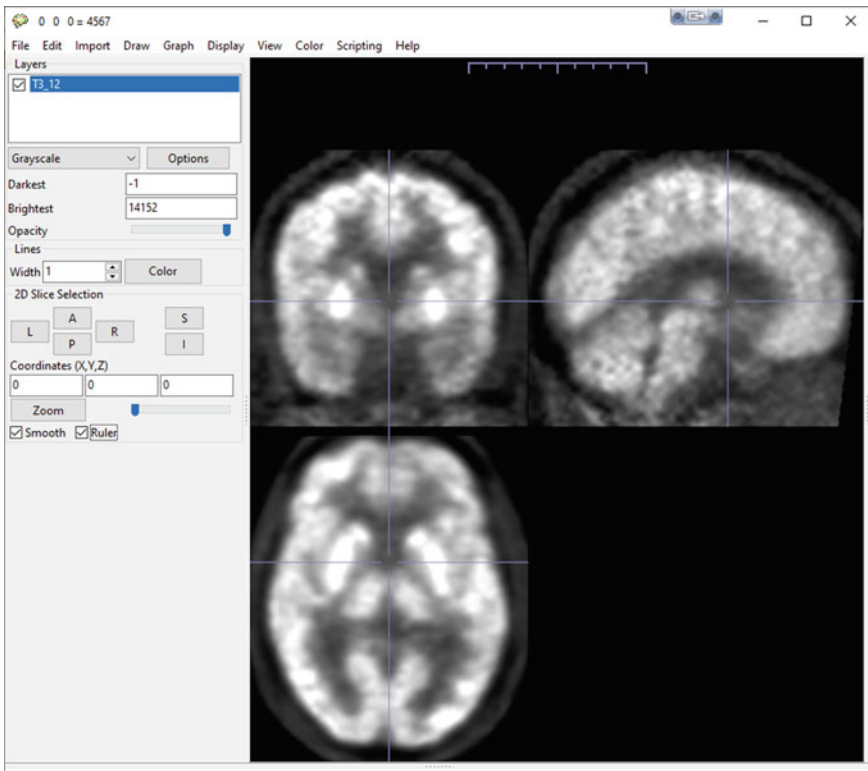


Fig. 2 Set: Darkset = -1; Brightest = 14,152 image PET to 12 months

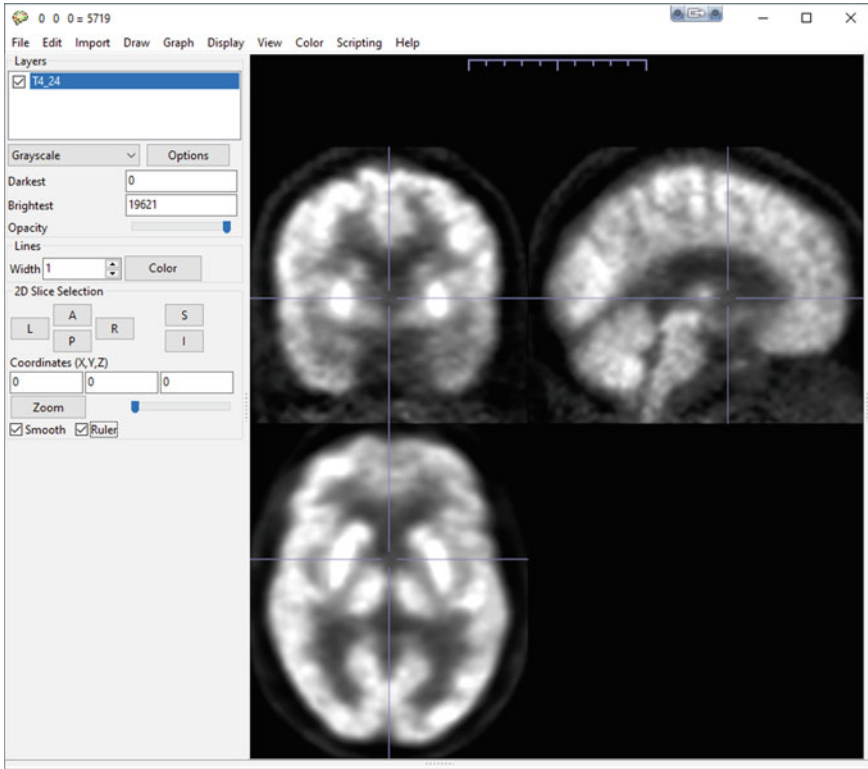


Fig. 3 Set: Darkset = 0; Brightest = 196,221 image PET to 24 months

evaluate an objective interpretation of the examined classes. Some of the numerical algorithms used are of different types such as spatial and connectivity algorithms widely used in the analysis of satellite images [19, 20].

- An algorithm or numerical index is a mathematical expression that transforms a set of data into a synthetic numerical value attributable to a feature of the analyzed image.
- Distance indices highlight in terms of space or time the distance between two or elements within an n-dimensional space.
- Connectivity indices express the degree of relationship between different objects spatially placed in different ways. Some shapes can express the relationships between different objects in numerical form, for example the relationship between area and perimeter allows you to distinguish the shapes of objects.
- The spatial indices evaluated with the depth of the pixel, i.e., with the value of radiance are able to describe the shape of the spots. The evaluation of the edges and the relationships found with the homogeneous areas are important information to evaluate.



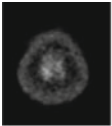
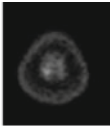
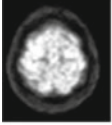
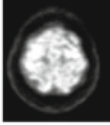
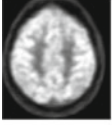
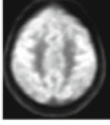
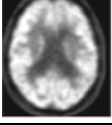
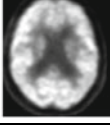
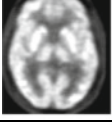
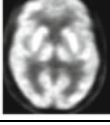
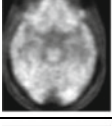
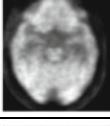
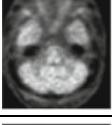
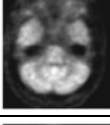
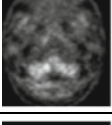
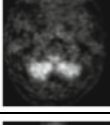
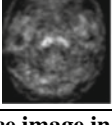
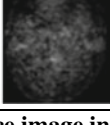
Top	Reference image in Z – T12	Reference image in Z – T24
80		
60		
40		
20		
0		
-20		
-40		
-60		
-80		
Down	Reference image in Z – T12	Reference image in Z – T24

Fig. 4 Reference images T12 and T24

## 4.1 Numerical Algorithms

Numerical algorithms are mathematical expressions capable of defining some information and synthesizing it in a numerical value. In fact, these algorithms measure the shape of objects, the relationship of objects, their abundance, and their spatio-temporal relationships. Where  $A_i$  is the abundance of object  $I$ .

$$A = \sum A_i \quad (1)$$

The Distance Index represents a fundamental element between two objects. The distance in Euclidean space relative to the Pythagorean theorem, given two points  $X'Y'$  and  $X''Y''$  their distance is given by the following:

$$d = \sqrt{(X' - X'')^2 + (Y' - Y'')^2} \quad (2)$$

the distance between two points is calculated as the shortest spatial interval, this is true for isotropic surfaces, if instead we have anisotropic surfaces the minimum distance between two points is generally not a straight line.

When we are in the presence of a group of points in a Euclidean space, for example objects that are scattered, the distance between the points can be measured for each element that we consider but we can also expect to measure a standard distance which represents the dispersion of the total. This distance can be calculated with the quadratic mean of the distance from the center of gravity.

$$d_1 = \sqrt{\left(\sum n/i = 1(d_{ic}^2/n)\right)} \quad (3)$$

$d_{ic}$  is the distance between each observation  $i$  is the middle center of all points  $c$ ,  $n$  the number of points.

## 4.2 Spatial Algorithms

For spatial algorithms we can indicate as a calculation procedure that describes the spatial characteristics of objects in a global system. These characteristics are both topological (dimensions, shape) and chorological (position relative to other objects of the same or different types). Spatial algorithms can describe the shape, evaluate the individual complexity of each grouping and that expressed collectively in a global system. Irregularity of the edges, the size of the area, interspersed and contact are fundamental parameters.

The shape of the elements are based on the difference between a geometric figure (circles, squares) and the interior of this figure, assuming maximum regularity for geometric figures in which the ratio between the suitably treated area and

the respective perimeter is approximately 1.

$$Y^1 = (2\sqrt{\pi A})/P \tag{4}$$

where A is the area and P the perimeter of the patch with  $g1 \simeq 1$  for circular patches and  $g1 < 1$  for non-circle or polygon-based shapes. Or simplified area perimeter report for elements represented in pixel and/or voxel format that we consider:

$$Y^3 = A/p^3 \tag{5}$$

The perimeter is calculated using the Sobel operator which is an algorithm used to perform edge detection.

The operator applies two  $3 \times 3$  kernels, ie two convolution matrices to the original image to compute the approximate values of the derivatives—one in the direction of the horizontal axis and one in the direction of the vertical axis.

If we call A the source image, and  $G_x$  and  $G_y$  the two images whose points respectively represent the approximate values of the horizontal and vertical derivatives, the operation is described by:

$$G_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} * A \text{ e } G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * A \tag{6}$$

### 4.3 Connectivity Indices

Connectivity algorithms express the degree of “relationship” between different objects placed in a spatial dimension. The graph elements are able to express in numerical form the chorological relationships between the different objects and at the same time to characterize their spatial schemes.

The connectivity algorithms express the maximum of the distances between the node i from each of the other nodes j. Any graph can be converted from a matrix where  $d_{ij} = 1$  if it exists.

$$K_i = \max d_{ij} \tag{7}$$

Accessibility index is given by the following:

$$A_1 = \sum_{i=1}^n d_{ij} \tag{8}$$

where  $d_{ij}$  is the number of nodes encountered to arrive at the chosen node. The value of the accessibility index is inversely proportional to the accessibility of the network node [21, 22].

### 4.4 Feature Selection

The first stage of processing a DICOM image from the ADNI repository involves reading the image and creating the matrix in a database. This matrix was subsequently elaborated through a feature selection through the numerical algorithms listed above, in order to memorize the value of the pixels to which the numerical values of some of their geometric properties have been associated, such as area, maximum dimensions  $\Delta x$  and  $\Delta y$  along the abscissa axis and along the ordinate axis, area of the rectangle  $\times \Delta x, \Delta y$ , moments of inertia  $J_x$  with respect to the abscissa axis and  $J_y$  with respect to the ordinate axis, coordinates of their barycenter  $G$ , etc. (see Fig. 5).

In this way, the images detected the following objects:

The PET/DICOM image is characterized by a matrix  $M$  with number rows  $m$  equal to the number of pixels and/or voxels and with a number of columns  $n$  equal to the number of geometric properties considered for each pixel. Below is a list of the fourteen descriptive variables which are the expression of the previously described numerical algorithms, the values of the feature selection as shown in Tables 1 and 2.

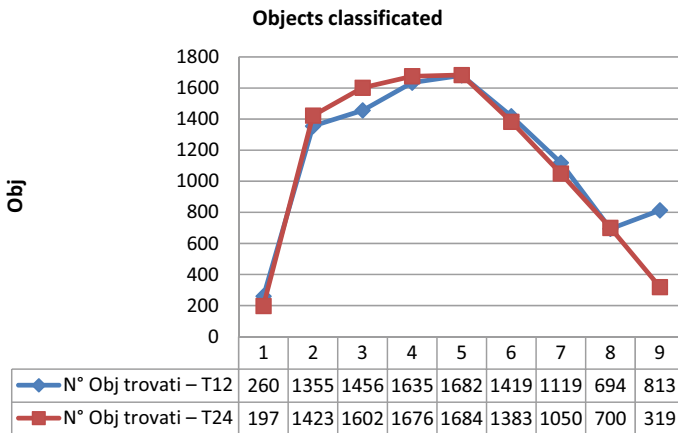


Fig. 5 Number of selected objects

**Table 1** Object considered

Z	N° Obj- T12	N° Obj- T24
80	260	197
60	1355	1423
40	1456	1602
20	1635	1676
0	1682	1684
-20	1419	1383
-40	1119	1050
-60	694	700
-80	813	319

**Table 2** Variables considered

Variable	Description
Nz	Z coordinate—Average depth
Nt	No of pixels of the object
Area	Attributes of the object: -Area
DeltaX	Xmx-Xmn
DeltaY	Ymx-Ymn
IdealArea	Ideal Area = DeltaX * Delta Y
Gx	• Barycentre X
Gy	• Barycentre Y
Jx	• Moment of inertia with respect to the X axis
Jy	• Moment of inertia with respect to the Y axis
Rx	• Radius of inertia X
Ry	• Radius of inertia Y
AreaRect	• Area of the circumscribed rectangle
RapportAAR	• relationship between area and area of the circumscribed rectangle

### 4.5 Factorial Analysis (PCA)

Factor analysis was applied to the matrix of all objects that have the description of the 14 labeled variables. Factor analysis allows to “order” in a vector distribution of data in order to maximize the variance and, through this information, reduce the size of the problem, represent the same amount of information with less data, transform the input data in a that the covariance matrix of the output data is diagonal and therefore the components of the data are uncorrelated, in k dimensions  $Z = (Z1, Z2, \dots, Zk)$  in terms of k variables  $Y1, Y2, \dots, Yk$ , linear combinations of the  $Z_j$ . It has:

$$Y_i = \sum_j b_{ij} Z_j \quad (i = 1, 2, \dots, k) \tag{9}$$

where  $b_{ij}$  are constants to be determined.  $Y_i$  is called the main components of the variable  $Z$  and assuming they are not related to each other ordered by importance, in the explanation of the variability of  $Z$  we have:

$$\text{cov}(Y_i, Y_j) = 0 \quad (i \neq j) \tag{10}$$

$$V(Y_1) \geq V(Y_2) \geq \dots \geq V(Y_k) \tag{11}$$

where  $\text{cov}$  is covariance and  $V$  is variance. Without loss of generality, we can assume that the variables  $Z_i$  are standardized, with mean equal to 0 and variance equal to 1, so as to eliminate the influence of the origin and the unit of measurement data, so that it results the following expression:

$$Z_j = (X_j - \mu_j) / \sigma_j \tag{12}$$

Also, impose the condition that the overall variance of  $Z_j$  is equal to that of  $Y_i$ , i.e.:

$$\sum_i V(Y_i) = \sum_i V(Z_i) = k \tag{13}$$

At last, suppose that the vectors

$$b_i = (b_{i,1}, b_{i,2}, \dots, b_{i,k}) \tag{14}$$

have unit length, i.e., they fulfill the condition:

$$\sum_j b_{ij}^2 = 1 \quad (i = 1, 2, \dots, k) \tag{15}$$

On account of this, the vectors  $b_i$  that maximize the variance of  $Y_1$ , of  $Y_2$ , ..., to  $Y_k$  with the constraints (3) and (4), are the eigenvectors of the matrix  $C$  of the coefficients of correlation between the variables  $Z_j$ , which correspond to the eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_k$  of  $C$ , sorted by non-increasing value. We then have:

$$|C - \lambda I| = 0 \tag{16}$$

$$b_i(C - \lambda_i I) = 0 \tag{17}$$

where  $I$  is the unit matrix. The matrix  $C$  is symmetric and positive definite for which the solutions  $\lambda_i$  of the (8) are non-negative and such that their sum (trace of the matrix  $C$ ) is equal to  $k$ . We then have:

**Table 3** Variance values explained on cases 10,433—T12

F1	F2	F3	F4	F5
0.42	0.68	0.79	0.86	0.92

$$\sum_i \lambda_i = k \quad (i = 1, 2, \dots, k) \tag{18}$$

The variance of the *i*-th component is:

$$V(Y_i) = \lambda_i \tag{19}$$

And the contribution of *Y<sub>i</sub>* to the overall variance is:

$$P_i = V(Y_i)/k = \lambda_i/k \tag{20}$$

Tables 3 and 5 shows the variance values explained according to the main components extracted.

Tables 4 and 6 shows the weight measurements of the variables obtained on each factor.

**Table 4** Weight measurements T12

Variable	F1	F2	F3	F4	F5
Nzm—F5	-0.1884	-0.0368	0.1249	-0.0119	0.9636
Nt—F2	0.1068	0.99	0.0072	0.0419	-0.0097
Area—F2	0.1068	0.99	0.0072	0.0419	-0.0097
DeltaX—F1	0.9666	0.0591	0.1026	-0.0153	-0.014
DeltaY—F1	0.9504	0.0884	0.0133	-0.0128	0.0124
ArIdl—F1	0.9081	0.13	0.0666	-0.0056	-0.0288
GX—F3	0.1451	0.0238	0.8552	-0.0152	-0.0445
Gy—F3	0.0896	-0.0043	0.8377	0.0588	0.1512
Jx—F2	0.0646	0.9935	0.0099	0.0283	-0.0217
Jy—F2	0.0651	0.9934	0.0099	0.0283	-0.0217
Rx—F1	0.9644	0.0482	0.0981	-0.012	-0.083
Ry—F1	0.9531	0.0398	0.0822	-0.0164	-0.0932
ArRe—F1	0.8987	0.0792	0.0925	-0.0029	-0.1664
R/AAR—F4	-0.0368	0.0919	-0.0328	0.9937	-0.0085

**Table 5** Variance values explained on cases 10,034—T24

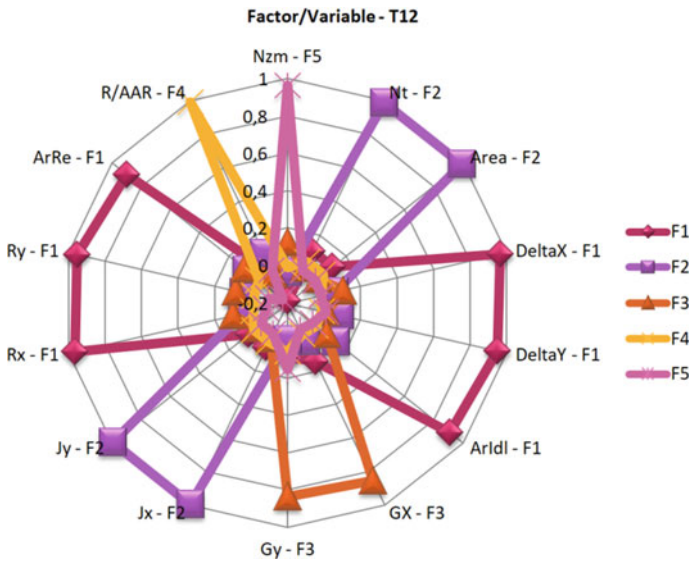
F1	F2	F3	F4	F5
0.41	0.68	0.78	0.85	0.92

**Table 6** Weight measurements T24

Variable	F1	F2	F3	F4	F5
Nzm—F5	-0.2181	-0.0459	0.0574	0.0032	0.9528
Nt—F2	0.1075	0.9899	0.004	0.06	-0.0109
Area—F2	0.1075	0.9899	0.004	0.06	-0.0109
DeltaX—F1	0.9679	0.065	0.084	-0.0175	-0.0052
DeltaY—F1	0.9446	0.1025	-0.0202	-0.0099	0.0377
ArIdl—F1	0.8952	0.1536	0.0331	-0.0089	0.0098
GX—F3	0.1016	0.0245	0.8632	-0.0279	-0.0908
Gy—F3	0.066	-0.002	0.8617	0.0097	0.1404
Jx—F2	0.0629	0.994	0.0142	0.0199	-0.0204
Jy—F2	0.0633	0.994	0.0142	0.0199	-0.0204
Rx—F1	0.9561	0.0357	0.0849	-0.0163	-0.1312
Ry—F1	0.9417	0.0254	0.068	-0.0195	-0.1506
ArRe—F1	0.8821	0.057	0.0804	-0.0058	-0.2254
R/AAR—F4	-0.0408	0.1045	-0.0026	0.9935	0.0024

### 4.6 Results of the Proposed Methodology

The results of the distribution of the variables in a reduced multidimensional space show that the factor 5 in Figs. 6 and 7 has a homogeneous distribution in all the



**Fig. 6** Variables in a multi-dimensional space



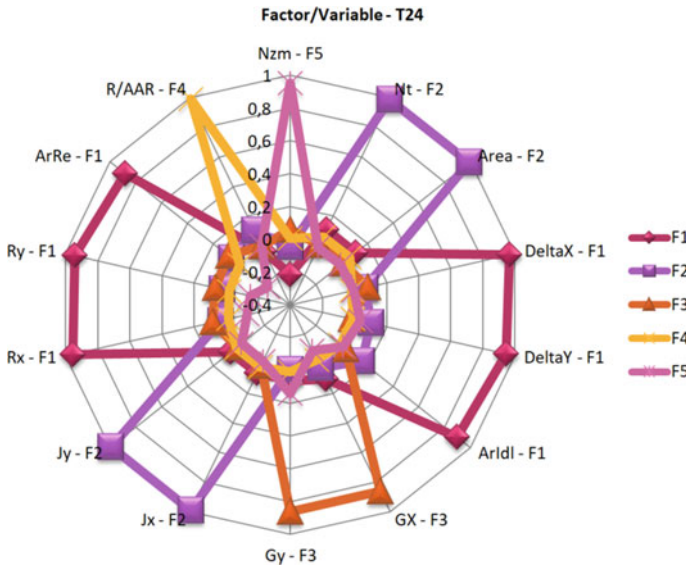


Fig. 7 Variables in a multi-dimensional space

explored variables with a preponderance in one variable that most characterizes this type of analysis, the most emerging variable is the depth in pixels expressed by the radiance of the elements belonging to the analyzed DICOM images.

Once the size of the definition space was reduced, I could reconstruct the images with the new values normalized with respect to the “feature selection”—greater than 1. This method required that the principal components from which the corresponding eigenvalues are less than 1 are not included in the reference model. This criterion is because the data are subject to a reduction of scale and therefore it can be assumed that the eigenvalue associated with each PC represents the number of variables whose variability is captured by the principal component. Therefore, if a PC does not represent at least one variable, it is not needed for model building. It is necessary to carefully consider the deviation of a PC whose eigenvalue is very close to 1, which can lead to not considering the variability explained by it, which could be non-negligible. Eigenvalues greater than 1 is a criterion also known as Kaiser’s criterion, you choose those components with an associated eigenvalue greater than 1.

The eigenvalue is a number that gives the variance explained by the component, since initially the variance explained by each individual variable is equal to 1, it would not make sense to take a component (which is a combination of variables) with variance less than 1, hence Kaiser’s rule.

A high eigenvalue corresponds to a higher variance, which is equivalent to returning the results of the reconstructed images these tables with decreasing values, so the former will always be associated with the most important factor.

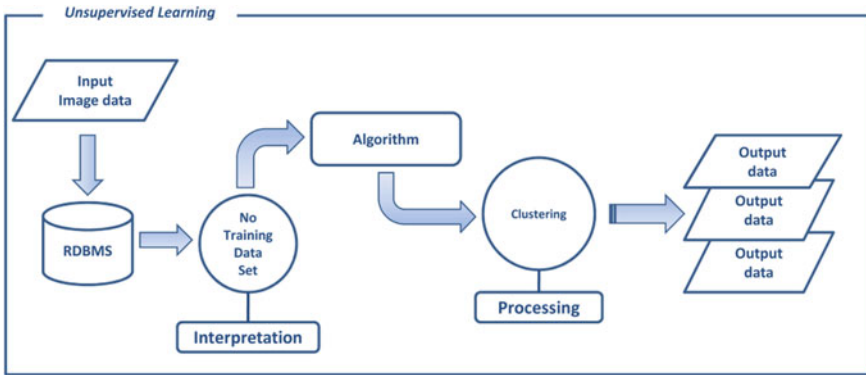


Fig. 8 Schema of methodology used

Figure 8 shows the outline of the methodology proposed for the study and analysis of biomedical images.

The created software runs on the Windows platform. Figure 9 shows the software GUI created after data extraction on the factor 5 (right) of the longitudinal image compared to its counterpart Z4 section (left) at time T24 of the DICOM data.

The results obtained in the factor F5 as shown in Fig. 10, show the maximum overall variance explained which is the sum of all the elements with maximum variance derived from the factor 1 to the factor 5 from the variables extracted with

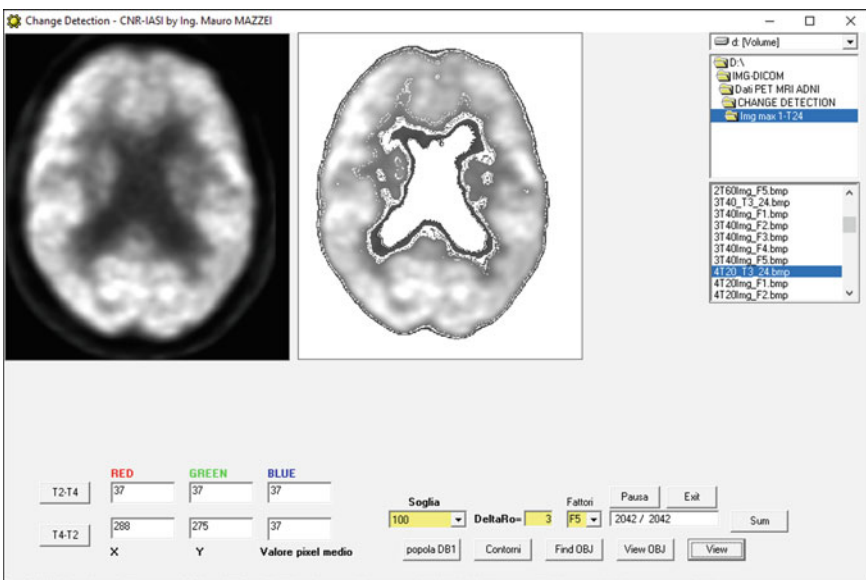


Fig. 9 GUI of software realized

Z	F1	F2	F3	F4	F5
80(1)					
60(2)					
40(3)					
20(4)					
0(5)					
20(6)					
40(7)					
60(8)					
80(9)					

**Fig. 10** Images reconstruction T12

the feature selection algorithms. These reconstructions are shown in Fig. 10 at time T12 and Fig. 11 and time T24 [23].

The graphically obtained results are shown in Fig. 12 for greater visual comparison of image reconstructions obtained by the proposed method.

Z	F1	F2	F3	F4	F5
80(1)					
60(2)					
40(3)					
20(4)					
0(5)					
20(6)					
40(7)					
60(8)					
80(9)					

Fig. 11 Images reconstruction T24

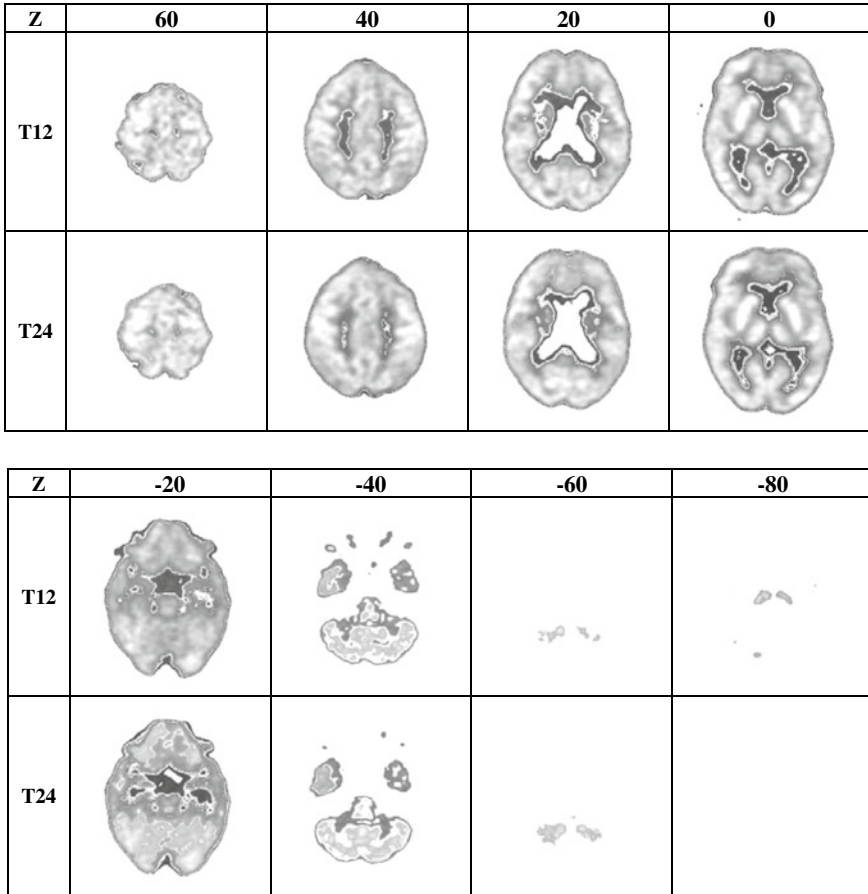


Fig. 12 Images reconstruction compared

## 5 Conclusion

The analysis of biomedical images through specific algorithms with the aid of calculation systems becomes a challenge to be taken up in the field of biomedical sciences. The continuous study and definition of specific Machine Learning algorithms will lead to a significant improvement in image processing which will become increasingly precise and detailed in view of scientific progress.

This work has its particularity in the classification of data identified by compact and well distinguishable clusters without having a knowledge of their nature; this type of classification is well defined as unsupervised, i.e., it expresses a truly objective evaluation of the entire set of data analyzed.

In the biomedical field, it is of paramount importance to obtain evidence for comparison as objective as possible so that there is no doubt about the interpretation

of data especially in medical diagnosis. In scientific disciplines, it is commonplace to refer to certain data; in this field, computational systems are very efficient in evaluating data and the complexity of variables to be compared, as opposed to a subjective interpretation by human beings.

This work proposes new scenarios in the evaluation and interpretation of biomedical images, this work aims to contribute to providing concrete help in the evaluation and diagnosis of biomedical images.

## References

1. Stauer, C., Grimson, W.E.L., Adaptive background mixture models for real-time tracking, Proceedings IEEE, Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, 1999.
2. Bruzzone, L., Prieto, D. F., Automatic analysis of the difference image for unsupervised change detection, IEEE Transactions on Geoscience and Remote Sensing, vol. 38, no. 3, pp. 1171–1182, May 2000
3. Bruzzone, L., Prieto, D. F., An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images, IEEE Trans. Image Processing, vol. 11, no. 4, pp. 452–466, April 2002
4. Elgammal, A., Duraiswami, R., Harwood, D., Davis, L., Background and Foreground Modeling Using Non-parametric Kernel Density Estimation for Visual Surveillance, Proceedings of the IEEE, vol. 90, No. 7, 2002.
5. Bruckstein, A. M., Donoho, D. L., Elad, M., From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images, SIAM Review Society for Industrial and Applied Mathematics, Vol. 51, No. 1, pp. 34–81, 2009.
6. Donoho, D. L., Huo, X., Uncertainty principles and ideal atomic decomposition IEEE Transaction Information Theory, vol. 47, pp. 2845–2862, 2001
7. Candes, E. J., Romberg, J., Tao, T.: Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information. IEEE Transactions on Information Theory, vol. 52, no. 2. pp. 489–509, 2006.
8. Patriarche, J.W., Erickson, B.J. A Review of the Automated Detection of Change in Serial Imaging Studies of the Brain, J. Digital Imaging, 17(3): 158–174, 2004.
9. Image Change Detection Algorithms: a Systematic Survey. IEEE Trans Image Process. Vol. 14, 2005
10. Automatic Change Detection in Multi-Modal Serial MRI: Application to MultipleSclerosis Lesion Evolution, Neuroimage, Vol. 20, 2003.
11. Candés, E. J., Romberg, J., Tao, T.: Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information. IEEE Transactionson Information Theory, vol. 52, no. 2. pp. 489–509, 2006.
12. Candés, E., Romberg, J., Sparsity and Incoherence In Compressive Sampling. J.Inverse Problems. vol. 23 no. 3, pp. 969–985, 2007.
13. Candés, E., Li, Xiaodong., Ma, Yi., Wright, J., Robust Principal Component analysis. Journal of the ACM, vol. 58, no. 3, 2011.
14. Aharon, M., Elad, M., Bruckstein A., K-SVD: an algorithm for designing over complete dictionaries for sparse representation, IEEE Transactions on Signal Processing, vol. 54, no. 11, pp. 4311–4322, 2006.
15. Nguyen, L.H., Tran, T.D: A Sparsity Driven Joint Image Registration And Change Detection Technique For SAR Imagery. ICASSP, pp. 2798–2801, 2010.
16. Turk, M., Pentland, A.: Eigenfaces for Recoregnition. J. Cognitive Neuroscience. vol. 3, no. 1, pp. 71–86, 1991.

17. Mazzei, M. (2021). An Unsupervised Machine Learning Approach for Medical Image Analysis. In: Arai, K. (eds) *Advances in Information and Communication. FICC 2021. Advances in Intelligent Systems and Computing*, vol 1364. Springer, Cham. [https://doi.org/10.1007/978-3-030-73103-8\\_58](https://doi.org/10.1007/978-3-030-73103-8_58)
18. Ambrose J., computerized transverse axial scanning: Part 2., Clinical application, 1973.
19. M. Mazzei and A. L. Palma, "Spatial Statistical Models for the Evaluation of the Landscape," in *Computational Science and Its Applications—ICCSA 2013*, 2013, pp. 419–432.
20. M. Mazzei and A. L. Palma, "Evaluating principal components analysis of particular spatial statistical models," in *Sixth International Conference on Advanced Geographic Information Systems, Applications, and Services*, 2014, pp. 24–30.
21. M. Mazzei, "Software development for unsupervised approach to identification of a multi temporal spatial analysis model," in *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (ICCV)*, 2018, pp. 85–91.
22. M. Mazzei, "An Unsupervised Machine Learning Approach in Remote Sensing Data," *Computational Science and Its Applications—ICCSA 2019*. pp. 435–447, 2019
23. Kirsten Boedeker, PhD, DABR, AiCE Deep Learning Reconstruction: Bringing the power of Ultra-High Resolution CT to routine imaging, Canon Medical System Corporation, 2019.
24. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y., Robust Face Recognition Via Sparse Representation. *Transactions on Pattern Analysis and Machine Intelligence, IEEE*, vol. 31, no. 2, pp. 210–227, 2009.
25. Brooks R.A., Theory of image reconstruction in computed tomography, Radiology, 1975.
26. Dale L Bailey, David W Townsend, Peter E Valk, and Michael N Maisey. *Positron Emission Tomography*. Springer-Verlag, London, 2005.
27. Dalla Palma L., Pozzi-Mucelli R.S., Image quality criteria for computed tomography, Report 20, 1989.
28. R.C Gonzales, R.E. Woods, *Digital Image Processing*. New Jersey: Prentice Hall, 2008
29. Joseph P.M., *Artefacts in computed tomography*, St. Louis Mosby, 1981.
30. LeCun Y., L. Bottou, Y. Bengio, and P. Haffner., Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, november 1998
31. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, 2011.
32. Pratt W. K., "Digital image processing" , John Wiley & Sons Ed., New York, 1978
33. Katzer, M., Kummert, F., Sagerer, G.: *Methods for Automatic Microarray Image Segmentation*. *IEEE Transactions on NanoBioscience* 2(4), 202–214 (2003)
34. Pappas, T.N.: *An Adaptive Clustering Algorithm for Image Segmentation*. *IEEE Transactions on Signal Processing* 40(4), 901–914 (1992)
35. Yonghong, H., Englehart, K.B., Hudgins, B., Chan, A.D.C.: *A Gaussian Mixture Model Based Classification Scheme for Myoelectric Control of Powered Upper Limb Prostheses*. *IEEE Transactions on Biomedical Engineering* 52(11), 1801–1811 (2005)
36. Bouguila, N., Ziou, D., Monga, E.: *Practical Bayesian Estimation of a Finite Beta Mixture Through Gibbs Sampling and its Applications*. *Statistics and Computing* 16(2), 215–225 (2006)
37. Varvara Nika, *Machine Learning and its application in automatic change detection in medical images*, 2014.
38. Andrés García-Florian, Ángel Ferreira-Santiago, Oscar Camacho Nieto, Cornelio Yáñez-Márquez: *A machine learning approach to medical image classification: Detecting age-related macular degeneration in fundus images*. *Comput. Electr. Eng.* 75: 218–229 (2019)
39. Xing-jiang Yang, Yong Zhou, Qingxing Zhu, Zhendong Wu: *Joint graph regularized extreme learning machine for multi-label image classification*. *J. Comput. Meth. in Science and Engineering* 18(1): 213–219 (2018)
40. Perikumar Javia, Aman Rana, Nathan Shapiro, Pratik Shah: *Machine Learning Algorithms for Classification of Microcirculation Images from Septic and Non-Septic Patients*. *CoRR abs/1811.02659* (2018)

# U-Net##: A Powerful Novel Architecture for Medical Image Segmentation



Firat Korkmaz 

**Abstract** As medical image segmentation has been one of the most widely implemented tasks in deep learning, there have been various solutions proposed for its applications to achieve better results. The encoder-decoder based U-Net architecture and its variants have shown outstanding performance in this field. However, most of these solutions have limited capacity to extract sufficient features from the input images. In this paper, we propose a powerful novel architecture named U-Net##, which consists of multiple overlapping U-Net pathways and has the strategies of sharing feature maps between parallel neural networks, using auxiliary convolutional blocks for additional feature extractions and deep supervision, so that it performs as a boosted U-Net model for medical image segmentation. Our architecture is essentially a combination of encoder-decoder based multiple U-Net pathways which have different depth levels, and all their overlapping feature maps on the same sampling steps share their own feature data with the others by following a specific addition rule. Each network pathway has its own concatenated long skip connections from their encoder to decoder sections, and the final output is obtained with deep supervision method. All these strategies help the model explore much more features effectively and achieve higher accuracy. PyTorch implementation of the U-Net## with step-by-step coding is available here: <https://github.com/firatkorkmaz/unetsharp>

**Keywords** Parallel neural networks · Auxiliary convolutional blocks · Deep supervision · Medical image segmentation

## 1 Introduction

Since medical image segmentation has been playing a significant role in biomedical diagnostics, there have been many studies in deep learning to develop effective methods for achieving satisfactory segmentation results. The most common proposed

---

F. Korkmaz (✉)  
Istanbul Technical University, Istanbul, Turkey  
e-mail: [korkmazfi@itu.edu.tr](mailto:korkmazfi@itu.edu.tr)



models are variants of the encoder-decoder based convolutional neural networks such as Fully Convolutional Networks (FCNs) [1], DeepLab network versions [2] and the U-Net [3] architecture which is widely used in medical image segmentation due to performing quite well in this task and segmenting images effectively even if there is limited amount of data. As the feature maps from different scales explore different levels of information, such models with encoder-decoder based structures use skip connections which combine low-level feature maps from the encoder network with high-level feature maps from the decoder network, so that the fine-grained details on the images are preserved during the training process. While the Fully Convolutional Networks (FCNs) use added skip connections [4] to preserve image resolutions throughout the network and have few parameters due to using  $1 \times 1$  convolution filters, the U-Net architecture uses  $3 \times 3$  convolution filters and concatenated skip connections which are followed by dense layers, resulting in better segmentation. And ever since then, there have been various proposed models based on the U-Net architecture.

After He et al. [5] introduced deep residual networks with identity mappings, Zhang et al. [6] combined this method with U-Net and proposed the ResU-Net architecture. Then, Oktay et al. [7] introduced Attention U-Net which was based on the integration of attention gates with the U-Net architecture, resulting in the network's attention to the significant regions in feature maps. Jha et al. [8] improved the ResU-Net architecture and proposed ResU-Net++ by adding Squeeze and Excitation [9] blocks between the down-sampled convolutional blocks, using Attention Gates [7] before the skip connections and placing Astrous Spatial Pyramid Pooling (ASPP) [10] module between encoder and decoder networks as a bridge. Despite being proposed for image classification tasks in the first place, another effective strategy for achieving prediction results with higher accuracy is Deep Supervision [11] which is also applicable to the architectures for image segmentation. This method is based on additionally supervising the intermediate layers of a convolutional neural network and directly involving them in the loss evaluation process, rather than using only the final output layer for this operation. So that, all these supervised hidden layers collectively participate in the learning process of the model. As shown in [12], this method also helps preventing the vanishing gradients problem [13] in training deep neural networks.

Based on reducing the semantic difference between the feature maps of encoder and decoder networks, Zhou et al. [14] proposed U-Net++ architecture to capture finer details from medical images by using nested and dense skip connections. Additionally, it also uses Deep Supervision to obtain results with higher accuracy. Later, Huang et al. [15] proposed U-Net3+ by placing full-scale skip connections between the feature maps from different scales and this network also uses Deep Supervision. Another proposed model for medical image segmentation which contains the U-Net architecture as a part of itself is KiU-Net [16]. This model has two network pathways which are Kite-Net and U-Net, working together in a parallel way. On the same sampling steps, the feature maps from both the networks are added to the corresponding ones from their opposite networks, where auxiliary convolution blocks are

applied to the added feature maps just before changing their spatial sizes for the addition.

By using multiple parallel neural networks, auxiliary convolutional blocks and deep supervision method, we propose a powerful novel architecture named U-Net## to address the demand for achieving more efficient predictions with higher accuracy and gaining faster network convergence in medical image segmentation. Our model uses overlapping multiple U-Net pathways with different depth levels, where all these U-Nets in the encoding steps gradually branch from a ceiling network that has constant-size layers. And on the decoding steps, they are gradually merged back to this ceiling network by concatenation. All these network branches work together in a parallel way while the corresponding feature maps on the same sampling steps share their own feature data with the others by a specific rule. The U-Net## model is evaluated on the TCIA-LGG Segmentation Dataset, which was obtained from The Cancer Imaging Archive (TCIA) and manually annotated by Buda et al. [17], to segment the brain regions with FLAIR abnormalities on the related brain MRI images. This dataset is available here: <https://www.kaggle.com/datasets/mateuszbeda/lgg-mri-segmentation>

## 2 The Proposed Architecture

Our architecture contains overlapping multiple U-Net pathways with different depth levels which work together in correlation with each other. Figure 1 shows the details of the architecture with the connections between feature maps from different network pathways which are passed through auxiliary convolutional blocks before their addition for more efficient feature extraction. As it is seen, the U-Net## architecture has multiple encoder networks and multiple decoder networks which belong to different U-Net pathways and this makes it act as a boosted U-Net architecture. By also taking the advantage of deep supervision strategy, the model provides better results with much higher accuracy.

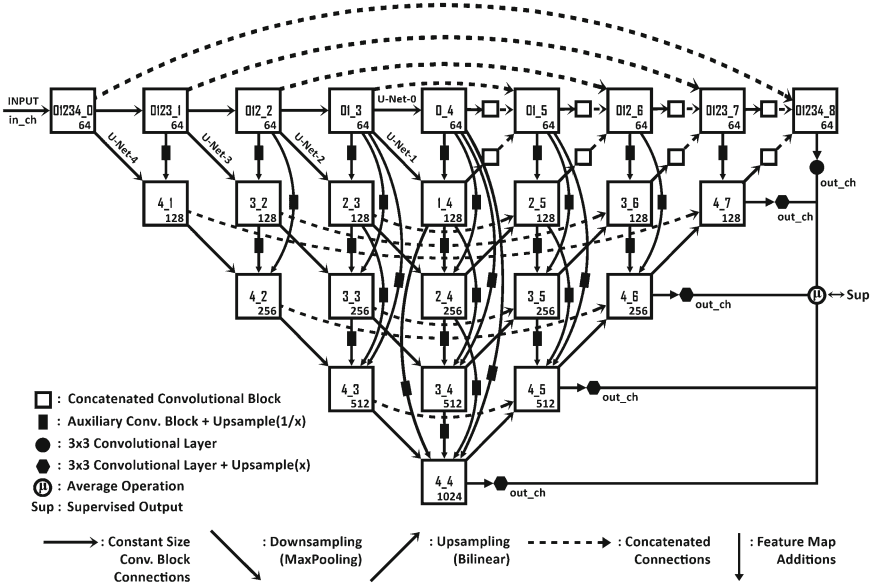
To understand the addition rule of the feature maps, the blocks in 3rd encoding step could be represented as:

$$B4\_3 = B4\_3 + AU_{(0.5)}(B3\_3) + AU_{(0.25)}(B2\_3) + AU_{(0.125)}(B01\_3) \quad (1)$$

$$B3\_3 = B3\_3 + AU_{(0.5)}(B2\_3) + AU_{(0.25)}(B01\_3) \quad (2)$$

$$B2\_3 = B2\_3 + AU_{(0.5)}(B01\_3) \quad (3)$$

Here, for the  $AU_{(scale)}(Bx)$  function,  $Bx$  is the identity of its input block and  $AU_{(scale)}$  represents that, firstly an Auxiliary convolutional block is applied to the output of its input block, and then a bilinear Upsample layer is applied to the output of this



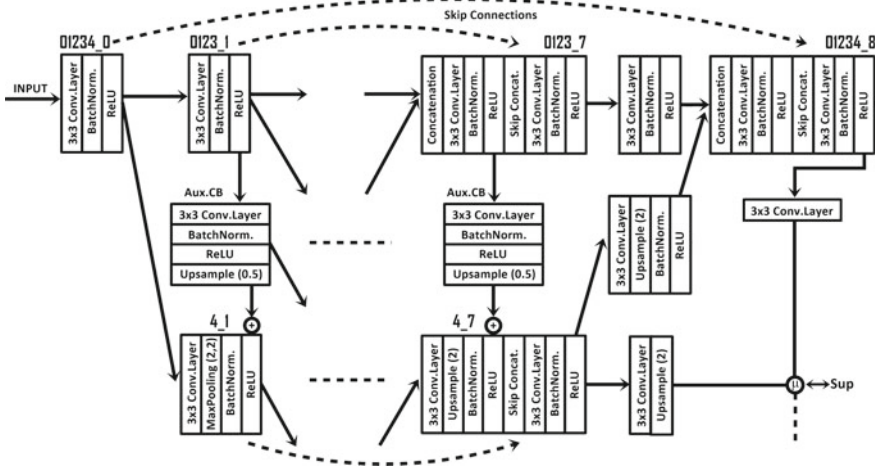
**Fig. 1** The block diagram of the U-Net## architecture with the overall view of its feature data transmission routes

Auxiliary convolutional block, where the up-sampling is done by the specified scale value. Also, the order of additions shows that one addition should not affect another.

### 2.1 Parallel Neural Networks

The U-Net## model starts with an initial block which has the same spatial size with the input images. This block is also the first block of all the parallel U-Net pathways. As for the notations of the block numbers, they show which network they belong to and what index of convolutional block sequence they have in these networks. For example, in the block of 0123\_1, the number 0123\_ says that this block belongs to U-Net-0 (the ceiling network), U-Net-1, U-Net-2 and U-Net-3, and the number \_1 says that this is the second block in all these four networks with the index = 1 value. Each U-Net has different encoding depth levels which vary between 1 and 4. The uppermost pathway is the ceiling network with constant spatial size layers and it does not have any down-sampling process. All the remaining encoder-decoder based networks below that are the U-Nets which branch one by one from this ceiling network while the encoding steps take place.

As it is seen in Fig. 2, every sampling step in the whole network is implemented with only one convolutional block which consists of one Convolutional layer, one Batch Normalization layer and one Rectifier Linear Unit (ReLU) layer. According to



**Fig. 2** The details of the convolutional blocks from the beginning and the end of the U-Net## architecture

the step type, there is also one additional Max Pooling layer if it is a down-sampling step or one bilinear Upsample layer with scale factor 2 if it is an up-sampling step. While each U-Net pathway is branched from the ceiling network on the encoder side, they are also gradually merged back to that ceiling network on the decoder side by concatenation. Each concatenation process to merge the U-Nets back is followed by transferring their output feature maps downwards for their additions, and then long skip connections are concatenated which are followed by one single convolutional block to decrease the filter numbers back to their previous state. Finally, all these network branches meet and end on a final block that has the 01234\_8 identity number.

## 2.2 Auxiliary Convolutional Blocks

When the encoder section starts, each U-Net branches from the ceiling network. However, to make all these U-Nets work together in correlation, they need to share their feature data with the other feature maps progressively. And rather than directly adding the feature maps from different U-Nets with each other, passing a feature map through an extra convolutional block before its addition provides much more efficient feature extraction. So, as seen in Fig. 2, auxiliary convolutional blocks are applied to the feature maps which are added to the other feature maps from different U-Net branches on the same encoding or decoding steps. An auxiliary convolutional block simply consists of one Convolutional layer with  $3 \times 3$  filters, one Batch Normalization layer and one Rectified Linear Unit (ReLU) layer. This block comes just before the necessary bilinear Upsample layer that is used to equalize the spatial size of the layers for their addition process. Applying the auxiliary convolutional blocks to

the feature maps for this purpose results in better overall network performance with higher accuracy.

As illustrated in Fig. 1, the rule of addition for the feature maps from parallel U-Nets says that each feature map of the U-Nets below the ceiling network collects all the other feature maps above themselves on the same encoding or decoding steps. Which means that all these additions are implemented downwards and all the feature data from different scales is always transmitted to the lower network branches which have higher depth levels. As a result, each U-Net also works as a feature collector and the undermost U-Net with the highest depth level collects the most diverse feature data from the upper networks, so that it becomes ready to be deeply supervised for the final output.

### 2.3 Deep Supervision

Rather than using only the final layer of the model where all the network branches meet and end, involving the hidden layers in the evaluation of loss function results in much better performance for the U-Net## architecture. As the overlapping U-Nets with different depth levels always share their own feature data with the other U-Net branches below themselves, the intermediate layers of the undermost U-Net branch collect the most diverse feature data from the other parallel U-Nets. Here, the outputs of the bottleneck block and the up-sampled decoder blocks in the undermost U-Net branch with the highest depth level are used together for the loss evaluation by passing each of these outputs through additional  $3 \times 3$  Convolutional layers and the necessary Upsample layers. Then, all the outputs that they produce are averaged and one single output layer is obtained. Finally, the Sigmoid activation function is applied to this single output layer and the result is used as the final output of the whole network for the loss evaluation. There is no necessity to use a specific loss function to evaluate the model, any loss function can be used with the U-Net## architecture.

## 3 Experiments and Results

### 3.1 Dataset and Pre-processing

The dataset used for training our proposed architecture is LGG Segmentation Dataset which contains the brain MRI images of 110 patients with FLAIR abnormality, and their corresponding segmentation masks. All the images have 3 channels and their binary image masks have 1 channel, provided in '.tif' format with the size of  $256 \times 256$  pixels and placed in specific folders for each patient. These folders include multiple brain MRI image slices and their masks in different numbers which vary from 20 to 88 among different patients. The total slice number from all the patients as

image-mask pairs is 3929 and this data is randomly divided into two sets, which are the training set with 3300 image-mask pairs, and the validation set with 629 image-mask pairs. An additional test set is not created to use more data in the training process. All the metrics are evaluated by using the validation set.

Before training, all the images and masks in both training and validation sets are resized to  $128 \times 128$  pixels. Also, all the images except the masks are normalized. For data augmentation; horizontal flip, vertical flip, random 90-degree rotation, zoom by random scale between 0.00 and 0.05, and additional random rotation between  $-20$  and  $20$  degrees are applied to both the images and the masks in only the training set with 0.5 probabilities.

### 3.2 Implementation Details

All the implementations were done on a Cloud GPU system with NVIDIA  $1 \times A100$  PCIE 40 GB GPU, by using the deep learning framework PyTorch. As the learning rate optimizer, Adam is chosen with the learning rate of  $1e-3$ . Also, the learning rate scheduler ReduceLROnPlateau is applied to decrease the learning rate if the evaluated loss value stops decreasing and does not improve for 10 epochs which is the patience value in the scheduler. Also, the reducing factor is set to 0.5 to divide the learning rate by 2 when it is necessary to reduce it, and the cooldown value is set to 5 which holds the scheduler from checking the improvement of learning rate for 5 epochs after a reducing process is implemented. For the training process, the epoch number is set to 100 and the batch number for data loader is fixed to 4 on training each model. Although there are other possible batch numbers which could give much better overall results for all the models, this number is set to 4 in this experiment deliberately, like reducing the spatial size of the dataset images from  $256 \times 256$  to  $128 \times 128$ , in order to observe the capacity of our model and to see what it can do while the other models cannot do in the given restrictive conditions and hyperparameters, throughout the 100 epochs of training process. The loss function which is used for all the models is Binary Cross Entropy Loss plus Dice Loss, named BCEDice Loss.

$$\mathcal{L}_{\text{bcedice}} = -\frac{1}{N} \sum_{i=1}^N [y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] + 1 - \frac{2 \sum_i (y_i \cdot p_i) + 1}{\sum_i y_i + \sum_i p_i + 1} \quad (4)$$

Additionally, Dice Coefficient and IOU values are also evaluated with the validation set for each trained model to compare their performances by these metrics.

$$\text{Dice} = \frac{2 \sum_i (y_i \cdot p_i) + 1}{\sum_i y_i + \sum_i p_i + 1} \quad (5)$$

$$IOU = \frac{\sum_i (y_i \cdot p_i) + 1}{\sum_i y_i + \sum_i p_i - \sum_i (y_i \cdot p_i) + 1} \tag{6}$$

### 3.3 Results

To verify and demonstrate the powerful performance of our proposed architecture, we compare our model with several state-of-the-art deep learning models including U-Net [3], AttU-Net [7], U-Net++ [14] and U-Net3+ [15]. As U-Net## uses deep supervision method, U-Net++ and U-Net3+ are also evaluated in their deeply supervised versions. The prediction results of each model after being trained with the mentioned hyperparameters and metrics for 100 epochs can be observed in Fig. 3 which shows that the U-Net## architecture gives the best results among all the evaluated models with the highest accuracy.

Here, a threshold of 0.3 is applied to the predicted images, so the pixel values which are equal to or higher than 0.3 are changed to 255, and the rest of the pixels with the values lower than 0.3 are changed to 0. Table 1 gives the comparison of performance results between the analyzed models in BCE-Dice Loss, Dice Coefficient and IOU (Jaccard) metrics which are evaluated with the validation set after training the models with the given hyperparameters. As we can see, U-Net## shows an outstanding performance and leaves all the other models behind with its capacity to learn from data quickly and efficiently, and hence makes the predictions much more effectively with the highest accuracy.

As U-Net3+ is coming after our U-Net## architecture in the results list with 2nd rank, Fig. 4 shows the change of Dice Coefficient values per epoch for both the models. This chart illustrates the power of U-Net## architecture and how it can

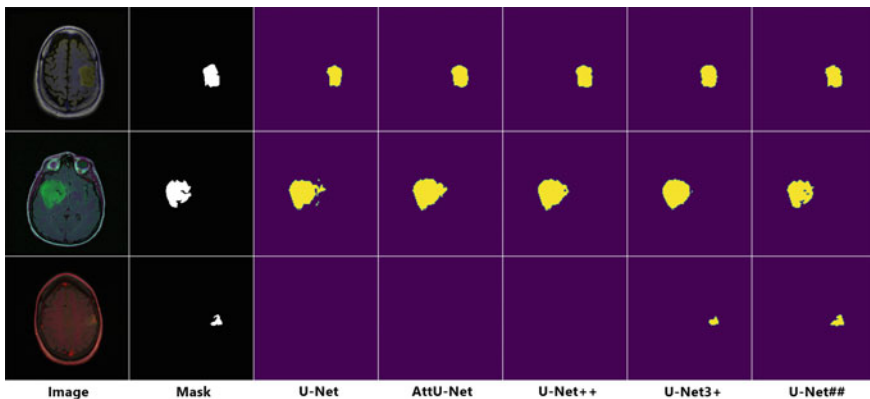
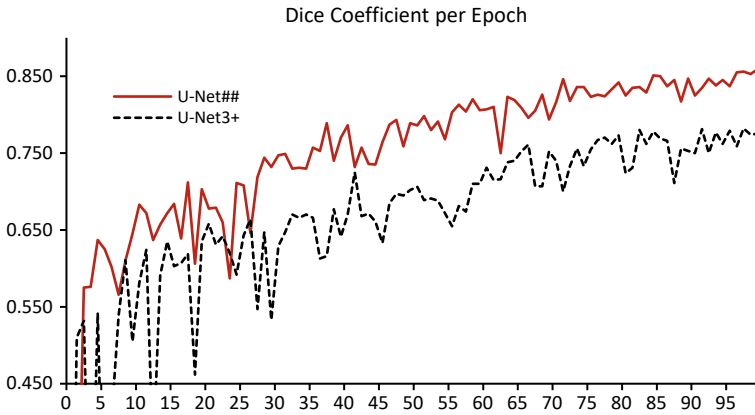


Fig. 3 Qualitative comparison between the prediction results of the models: U-Net, AttU-Net, U-Net++ , U-Net3+ and U-Net##

**Table 1** Results of the analyzed models by the specified metrics with validation data

Model	Parameters	BCE-Dice Loss	IOU (Jaccard)	Dice Coefficient
U-Net	34.52M	0.379	0.619	0.699
AttU-Net	34.87M	0.464	0.599	0.681
U-Net++	36.63M	0.394	0.639	0.702
U-Net3+	26.99M	0.351	0.703	0.774
<b>U-Net##</b>	43.21M	<b>0.165</b>	<b>0.792</b>	<b>0.859</b>



**Fig. 4** Comparison of the Dice Coefficient changes between U-Net3+ and U-Net## architectures during the training process

extract features from the medical images much more effectively. So, it is observed that even there are restrictive hyperparameters, U-Net## can achieve what many other models cannot do in a given training process.

## 4 Conclusions

To address the need for achieving more efficient and accurate medical image segmentations, we proposed a powerful novel architecture named U-Net##, with the strategies of sharing feature maps between encoder-decoder based parallel networks through auxiliary convolutional blocks and obtaining the output data with deep supervision. The most significant properties of this architecture are that it can learn faster from data efficiently and can do much more accurate segmentations even if there are restrictive hyperparameters or lower data qualities. Comparison of the U-Net## performance with the other state-of-the-art approaches shows that U-Net## has a great power and capacity to segment medical images much more effectively and accurately.



## References

1. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440. <https://doi.org/10.1109/CVPR.2015.7298965>
2. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834–848. IEEE, 2018. <https://doi.org/10.1109/TPAMI.2017.2699184>
3. 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](https://doi.org/10.1007/978-3-319-24574-4_28)
4. Drozdal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C.: The Importance of Skip Connections in Biomedical Image Segmentation. In: Deep Learning and Data Labeling for Medical Applications. DLMIA LABELS 2016 2016. LNCS, vol 10008, pp. 179–187. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46976-8\\_19](https://doi.org/10.1007/978-3-319-46976-8_19)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR), 2016, pp. 770–778. <https://doi.org/10.1109/CVPR.2016.90>
6. Zhang, Z., Liu, Q., Wang, Y.: Road Extraction by Deep Residual U-Net. In: IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 5, pp. 749–753. IEEE, 2018. <https://doi.org/10.1109/LGRS.2018.2802944>
7. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.J., Heinrich, M.P., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention U-Net: Learning Where to Look for the Pancreas. ArXiv, 2018. <https://arxiv.org/abs/1804.03999>
8. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., Lange, T.D., Halvorsen, P., Johansen, H.D.: ResUNet++: An Advanced Architecture for Medical Image Segmentation. 2019 IEEE International Symposium on Multimedia (ISM), 2019, pp. 225–2255. <https://doi.org/https://doi.org/10.1109/ISM46123.2019.00049>
9. Hu, J., Shen, L., Albanie, S.: Squeeze-and-excitation networks. In: Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141. <https://doi.org/10.1109/CVPR.2018.00745>
10. He, K., Zhang, X., Ren, S., Sun, J.: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8691. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-10578-9\\_23](https://doi.org/10.1007/978-3-319-10578-9_23)
11. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial Intelligence and Statistics, pp. 562–570. PMLR, 2015. <https://doi.org/10.48550/arXiv.1409.5185>
12. Wang, L., Lee, C.Y., Tu, Z., Lazebnik, S.: Training Deeper Convolutional Networks with Deep Supervision. ArXiv, 2015. <https://arxiv.org/abs/1505.02496>
13. Hochreiter, S.: The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. In: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998, vol. 6, pp. 107–116. <https://doi.org/10.1142/S0218488598000094>

14. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested u-net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-00889-5\\_1](https://doi.org/10.1007/978-3-030-00889-5_1)
15. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y., Wu, J.: UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1055–1059. IEEE, 2020. <https://doi.org/10.1109/ICASSP40776.2020.9053405>
16. Valanarasu, J.M.J., Sindagi, V.A., Hacıhalilolu, I., Patel, V.M.: KiU-Net: Towards Accurate Segmentation of Biomedical Images Using Over-Complete Representations. In: Medical Image Computing and Computer Assisted Intervention (MICCAI), 2020. LNCS, vol. 12264. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-59719-1\\_36](https://doi.org/10.1007/978-3-030-59719-1_36)
17. Buda, M., Saha, A., & Mazurowski, M.A.: Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. In: Computers in biology and medicine, 2019, vol. 109, pp. 218–225. <https://doi.org/10.1016/j.combiomed.2019.05.002>

# **Computer-Aided Detection/Diagnosis**

# Optimising Chest X-Rays for Image Analysis by Identifying and Removing Confounding Factors



**Shahab Aslani, Watjana Lilaonitkul, Vaishnavi Gnananathan, Divya Raj, Bojidar Rangelov, Alexandra L. Young, Yipeng Hu, Paul Taylor, Daniel C. Alexander, NCCID Collaborative, and Joseph Jacob**

**Abstract** During the COVID-19 pandemic, the sheer volume of imaging performed in an emergency setting for COVID-19 diagnosis has resulted in a wide variability of clinical CXR acquisitions. This variation is seen in the CXR projections used, image annotations added and in the inspiratory effort and degree of rotation of clinical images. The image analysis community has attempted to ease the burden on overstretched radiology departments during the pandemic by developing automated COVID-19 diagnostic algorithms, the input for which has been CXR imaging. Large publicly available CXR datasets have been leveraged to improve deep learning algorithms for COVID-19 diagnosis. Yet the variable quality of clinically-acquired CXRs within publicly available datasets could have a profound effect on algorithm performance. COVID-19 diagnosis may be inferred by an algorithm from non-anatomical features on an image such as image labels. These imaging shortcuts may be dataset-specific and limit the generalisability of AI systems. Understanding and correcting

---

S. Aslani · B. Rangelov · Y. Hu · D. C. Alexander · J. Jacob  
Centre for Medical Image Computing, University College London, London, UK

D. C. Alexander  
Department of Computer Science, University College London, London, UK

V. Gnananathan · D. Raj  
Department of Radiology, Royal Free London NHS Foundation Trust, London, UK

W. Lilaonitkul · P. Taylor  
Institute of Health Informatics, University College London, London, UK

S. Aslani (✉) · J. Jacob  
Department of Respiratory Medicine, University College London, London, UK  
e-mail: [a.shahab@ucl.ac.uk](mailto:a.shahab@ucl.ac.uk)

A. L. Young  
Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience,  
King's College London, London, UK

W. Lilaonitkul  
Health Data Research UK, London, UK

NCCID Collaborative  
London, UK

key potential biases in CXR images is therefore an essential first step prior to CXR image analysis. In this study, we propose a simple and effective step-wise approach to pre-processing a COVID-19 chest X-ray dataset to remove undesired biases. We perform ablation studies to show the impact of each individual step. The results suggest that using our proposed pipeline could increase accuracy of the baseline COVID-19 detection algorithm by up to 13%.

**Keywords** Computer-aided diagnosis · Chest X-ray · COVID-19 · Deep learning

## 1 Introduction

Medical research using artificial intelligence (AI) techniques applied to clinical data and imaging is transforming our understanding of health and disease. The application of machine learning and deep learning techniques has occurred alongside an exponential growth in healthcare data acquisition. The arrival of the SARS-COV2 virus and the response of healthcare teams around the world in data collection and data sharing exemplified the scale at which medical information can be collected today for clinical research purposes.

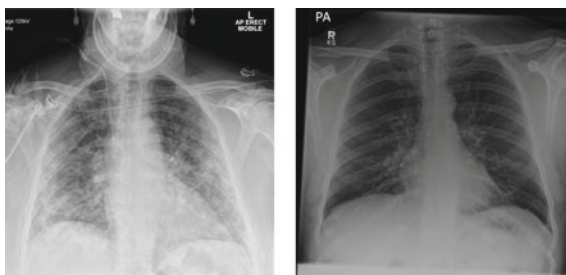
Large volumes of data are a prerequisite to train AI models. However, as datasets grow in size and complexity, it is important to be aware of biases that might be introduced into training AI datasets [9]. A major concern when using AI tools is that biases within training datasets may be propagated into deployed algorithms. Examples of biases include training datasets that are not representative of the target population but are imbalanced with regard to subject age, gender, ethnicity and socioeconomic or environmental factors. Such biases must be identified, understood and corrected to ensure that AI algorithms that could be used to assess patient health.

The SARS-COV2 virus has infected hundreds of millions of people across the world [11]. When the virus first emerged in late 2019, limited access to polymerase chain reaction testing kits led researchers to consider analysing medical images such as chest X-ray (CXR) and computed tomography (CT) scans to diagnose the disease and assess its severity [6, 18]. A hallmark of the image analysis approaches used was the reliance on deep learning systems to leverage the scale of data acquired in a timely way [10]. Over the past two years, many automated approaches have been proposed to detect COVID-19 infection using CXR images [1, 7, 10, 12, 14].

Though several of the proposed deep learning-based approaches have reported excellent discriminative performance, re-evaluation of several of these methods, particularly for CXR analyses has revealed that many were confounded by non-pathological features in the images such as human annotations [4]. These observations highlight the importance of having a suitable processing pipeline for CXR imaging that can remove confounding signals from the image, allowing the AI model to focus on detecting and quantifying pathologically important features alone.

With regard to CXR projection, the optimal CXR is captured with the patient standing upright and with the x-ray beam passing from the back of the patient to the

**Fig. 1** Anteroposterior (left) and Posteroanterior (right) chest X-ray images.



x-ray plate placed at the anterior aspect of their chest in a Posterior to Anterior (PA) acquisition. This allows the patient to perform the best expansion of their lungs in inspiration and enhances the detection of lung damage. As the heart is positioned in the anterior compartment of the chest, a PA radiograph provides a good approximation of the size of the heart relative to the size of the lungs and chest wall. A PA radiograph is the standard CXR acquisition, however if a patient cannot stand, a radiograph can be acquired with the patient sitting or lying down and placing the X-ray plate at the patient's back, known as an Anterior-Posterior (AP) projection. In an AP projection, the size of the heart relative to the chest is exaggerated as the x-ray beams pass through the patient from front to back. AP radiographs may also be associated with a smaller inspiratory volume. Figure 1 shows an example of CXR images acquired using PA and AP projections. All non-PA CXR acquisitions are labelled as such on the image using labels such as "AP", "sitting" or "supine". The type of CXR acquisition performed can provide information about the clinical status of a patient: a PA CXR would indicate that a patient was in better health than someone undergoing an AP CXR. COVID-19 imaging datasets typically contain a combination of AP and PA CXRs. Yet studies using AI tools to diagnose COVID-19 on CXRs invariably ignore the potential influence on algorithm performance that may result from the algorithm simply distinguishing AP versus PA radiographs, rather than actual lung disease features. The proportion of AP and PA CXRs also can vary across datasets, further biasing algorithms.

Other x-ray acquisitions can also be seen in publicly available COVID-19 chest radiographic datasets. Lateral view chest X-rays (Lat-CXR) are routinely performed alongside the frontal CXR acquisition in many countries such as the USA. However in the UK lateral CXRs are rarely performed in the acute setting. If a lateral view is performed in the UK it is often to confirm or refute the presence of a lesion suspected on the frontal CXR. Therefore there is a strong bias towards the presence of pathology in cases where a lateral CXR has been performed. Abdominal X-rays (AXR) are also not uncommonly seen in COVID-19 radiographic datasets alongside the frontal CXR acquisition, particularly in patients who presented with abdominal pain. Having AXRs and Lat-CXRs in COVID-19 datasets can introduce some biases and image noise as these images are not useful for COVID-19 detection using AI methods. Identifying and removing these images from a dataset can help AI models improve COVID-19 detection performance.

In this paper, we describe an automated pipeline for cleaning and pre-processing CXR images. Our pipeline can help standardize a CXR dataset for future analysis. The pipeline includes: defining the unique ID per patient and per patient's CXR session; removing noisy CXR images; identifying AP and PA CXR images; and applying a lung segmentation. We show that using the pipeline to clean a COVID-19 CXR dataset can improve the performance of the baseline models for COVID-19 detection. Moreover, this pipeline will improve CXR-based deep-learning models in other existing lung diseases and for analysis of CXR imaging in potential future pandemics. The open source code will be made available soon to anyone.

## 2 Materials and Methods

**Datasets.** In this work the UK National COVID-19 Chest Imaging Database (NCCID) [2, 3, 8] was analyzed. This dataset is a multi-center research database (comprising data from 14 NHS Hospital Trusts including 52 individual hospitals) comprising CXRs and CT scans from patients across the UK. The NCCID is a growing dataset initiated in January 2020 to enable the development of machine learning algorithms for the characterisation of patients hospitalized with COVID-19. All patients in NCCID have COVID-positive/-negative labels reflecting results of a SARS-CoV-2 RNA test via the Polymerase Chain Reaction (PCR) method. At the time of proposing our approach, 18,133 CXR images from 7629 patients were available for analysis.

**Method.** Since many centers were enrolled in NCCID, the imaging data, particularly the CXR imaging was highly heterogeneous with regard to imaging acquisitions. We propose an automated pipeline that can process a CXR dataset in a stepwise manner to create a standardized and homogeneous subset of the dataset, and limit potential source of bias. Our proposed pipeline includes:

- Assigning a unique ID for CXR images.
- Identifying and removing unnecessary/noisy X-ray images.
- Categorizing X-ray images into Posterior-Anterior and Anterior-Posterior acquisitions.
- Intensity normalization.
- Lung segmentation.

As a first step, we defined an image level identifier that encoded the anonymised patient identifier, the image acquisition date and the image acquisition time. This was to distinguish separate CXRs performed on an individual patient on the same day.

Within the DICOM folder of a CXR acquisition, it was not uncommon to encounter radiographs of the abdomen or lateral CXR projections. These are not suitable for COVID-19 detection. Therefore we identified them and omitted them from the dataset. Several DICOM tags were used to recognize these images automatically including 'Series Description', 'View Position', and 'Study Description'.

To identify AP and PA CXRs to allow for their separate processing we used the ‘Code Meaning’ sequence in the DICOM tag of the image. An example of AP and PA images can be seen in Fig. 1.

To maintain homogeneity and consistency in the dataset, all CXR images were scaled to a size of  $640 \times 512$ . A histogram equalization algorithm was then applied to increase the contrast within the images.

The majority of the CXR images in NCCID include text in their borders (outside of the lung area) which has been annotated manually by radiographers or radiologists. This type of information can result in undesired biases in AI models [4]. To remove these confounding signals, we used a lung segmentation algorithm that forced the model to concentrate on the lung region. A challenge in COVID-19 infected lungs is that they may contain peripherally placed areas of high density which can make the boundary between the lung and chest wall imperceptible. Accordingly, segmenting the lung in these cases can be a challenging task. To overcome this, we followed the idea proposed in [15]. The proposed segmentation architecture involves a variational encoder for data imputation, and a U-net shape network with encoder and decoder for segmentation. This model was specifically designed to segment lungs containing a high proportion of abnormalities including lung damaged by COVID-19 infection. Following successful lung extraction, we cropped the image to centralize the position of the lungs in the image.

### 3 Experimental Analysis and Results

In this section, we describe experiments to highlight the robustness of the pipeline for COVID-19 detection using the NCCID dataset. We designed a deep learning-based model to generate a binary classification for COVID-19 (COVID-positive/-negative) using the PCR-based COVID-19 classification as a ground truth. A 2-dimensional EfficientNet-B0 [17] pretrained on ImageNet [5] was selected for our baseline model. As a first experiment, we trained and tested our model using the raw NCCID dataset to which our proposed processing pipeline was not applied. 5-fold cross validation across the entire dataset was implemented (using 18,133 CXR images) and the average of the 5-fold over test set was reported as the overall performance metric. As can be seen in the Table 1, the AUC performance of the model was 72%.

In order to gain a better understanding of the robustness of the pipeline, an ablation study was performed. We repeated the above experiment several times, iteratively processing the dataset according to the steps mentioned in our processing pipeline. In this second experiment, those images contained within CXR DICOM folders that included the abdomen and lateral CXR views were omitted from the dataset. In total 1201 images were removed from the training and testing datasets. Our baseline model AUC performance on this processed dataset was 75% Table 1, a 3% increase compared to the unprocessed dataset.



**Table 1** Output performance of our baseline EfficientNet B0 [17] model for COVID-19 detection using different version of the NCCID dataset

Dataset	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUC
Raw data	0.65	0.56	0.74	0.68	0.61	0.72
AXRs and Lat-CXRs were removed	0.66	0.60	0.73	0.69	0.64	0.75
Only AP CXRs	0.73	0.74	0.73	0.73	0.73	0.81
Only PA CXRs	0.74	0.74	0.74	0.74	0.74	0.81
Processed AP CXRs	0.78	0.78	0.79	0.79	0.78	0.88
Processed PA CXRs	0.77	0.77	0.77	0.77	0.77	0.82

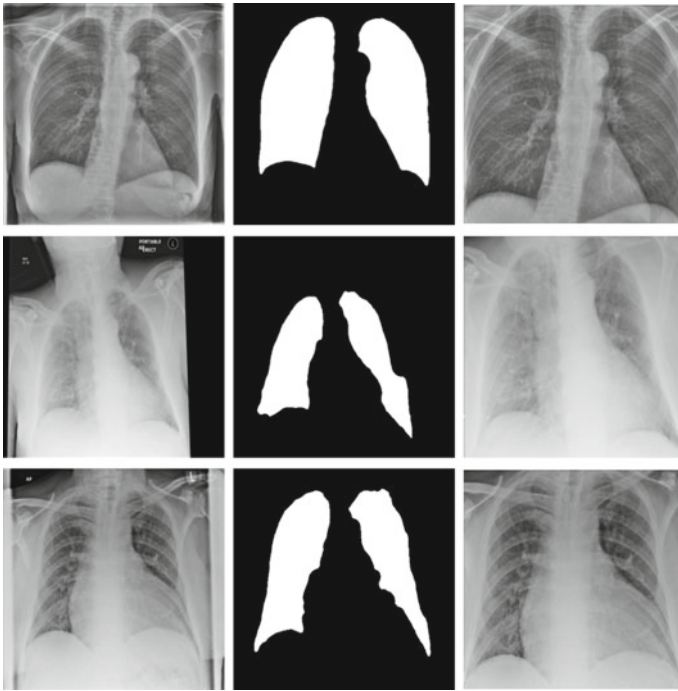
In the third experiment, using the aforementioned DICOM tags, we categorized the remaining images into two sub-groups (AP/PA). 71% of the images were assigned the AP label (12023 images). 29% were assigned the PA label (4909 images). We then repeated our cross validation experiments on AP and PA images separately. Table 1 shows the performance of our base-line model when regarding AP and PA images separately. When analysing AP CXRs, we have a 7% improvement in accuracy, and 6% improvement in AUC value.

To further process the dataset, we normalized the images using histogram equalization and used the model in [15] to segment the lungs. Then, we cropped the images centralizing the lungs by considering the maximum and minimum values for (x,y) pixel coordinates in the extracted mask. The output can be seen in the last column of the Fig. 2. In this experiment, we used the pre-processed AP and PA images as input to our baseline network. When using the CXRs with a segmented lung image Table 1, we obtained the highest AUC performance of our study: 88% and 82% for AP and PA images, respectively.

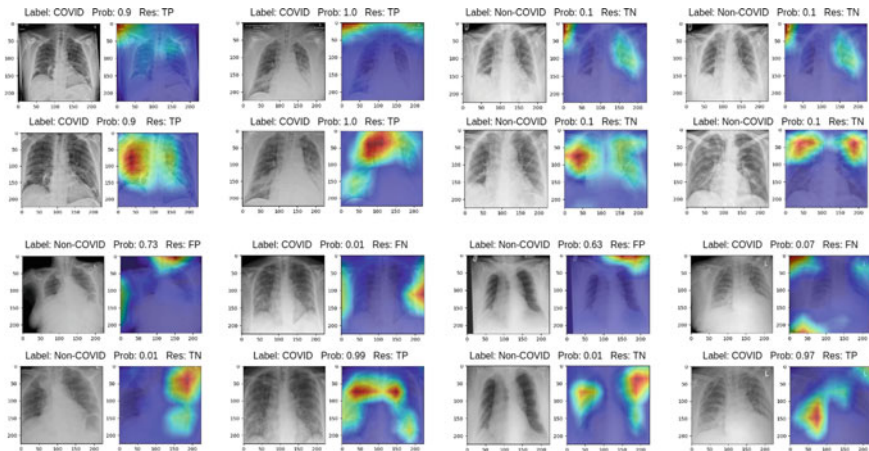
We implemented a saliency map using the GradCAM [16] algorithm to confirm whether the baseline model was making COVID-19 classification decisions using pathological signals from the lung. Figure 3 shows some examples of the baseline model output for COVID-19 detection including the saliency map visualization using both the original CXR image and the lung segmented version of the corresponding image. The results indicate that the processed CXR helps ensure the model is making decisions based on pathology rather than confounding annotations.

All experiments were performed in the python language<sup>1</sup> using PyTorch [13] on a Nvidia Titan RTX 24GB GPU. We trained our model using a stochastic gradient descent optimizer with an initial learning rate of 0.001, batch size of 32, and cross-

<sup>1</sup> <https://www.python.org>.



**Fig. 2** Lung segmentation process. We applied the lung segmentation algorithm on AP and PA X-ray images to obtain the lung masks which were used to extract the lungs from the original images (second column). We then cut the segmented images with centralizing the lung (third column)



**Fig. 3** Baseline model output for COVID-19 detection. The first and third rows show model performance on normal CXRs including a saliency map visualization. The second and fourth rows are the model outputs on processed CXRs of the respective first and third rows. At the top of each image, information includes: 'Label'=COVID-19 ground truth label, 'Prob'=output probability of the model for COVID-1, and 'Res'=TP/TN/FP/FN outcomes

entropy as the loss function. Data augmentation was applied on the training dataset using random rotation with a maximum of  $15^{\circ}\text{C}$ . In all experiments, the epoch number was set to 100. We used a modified version of EfficientNet including (B0, B1, B2, and etc.). We found that increasing the complexity of the model did not improve overall model performance and so EfficientNet-B0 was selected as our baseline model.

## 4 Discussion and Conclusion

We propose an automated multi-step processing pipeline to clean and standardize CXR datasets to minimise potential biases. The proposed pipeline improves dataset homogeneity and improves the performance of existing deep learning-based approaches to classify COVID-19 disease as exemplified on the UK NCCID dataset.

Defining unique identifiers for every single image in a dataset builds order into a complex dataset. Individual identifiers improve the certainty with which specific images are assigned into testing and training cohorts. When different models are then evaluated, the researcher can be confident that each model will be evaluated on identical data thereby producing a fair comparison of all available approaches.

In a large chest imaging dataset like NCCID, which continues to expand over time, images of different body parts as well as a variety of CXR acquisitions might be captured for specific purposes. Identifying the type of images captured and categorizing them appropriately can avoid misleading AI models. For instance, in the NCCID dataset, we identified lateral projection CXRs and abdominal CXRs and removed them from the dataset as these images are unhelpful for COVID-19 detection. We showed that by simply identifying and removing these unnecessary images, the AUC performance of our COVID-19 detection model improved by 3% Table 1.

As mentioned in Sect. 1, by separating the dataset into AP and PA CXRs, our goal was to avoid any possible undesired bias originating from the CXR acquisition in the baseline model. As shown in Table 1, training and testing a baseline model separately on AP and PA CXRs boosted the performance of the model with respect to all available measures including accuracy, sensitivity, specificity, precision, F1-score and AUC value.

The first row in Fig. 3, shows CXRs without any processing and their model output saliency maps. All the model output predictions were correct based on the PCR ground truth and the predicted output likelihood. However, when visualizing the saliency maps, it becomes obvious that the model is basing its predictions on non-pathological signals. Specifically, the strongest predictive signals originated from text in the corner of the images, calling into question the integrity of the model predictions. In the second row, comprising segmented lungs model prediction is unchanged but now focuses on pathological signals inside the lung.

When using deep-learning algorithms, model predictions influencing patient care may be based on imaging features unrelated to true biological damage, which could then result in patient harm. An image analysis should aim to confirm the biological plausibility of imaging features identified as having prognostic or diagnostic impor-

tance. The importance of these concepts are highlighted in the third row of Fig. 3, where on non-processed CXRs, the baseline model produces incorrect predictions by focusing on confounding signals outside of the lung. In the last row however, using the processed version of the images, allows the model to correctly predict outcomes using the pathological signals inside the lung.

Whilst this study was focused on evaluation of COVID-19 CXR imaging, we believe the steps in our CXR imaging pipeline will have relevance for all researchers attempting to analyse CXR imaging no matter what the underlying disease. We also believe that such pipelines will greatly aid analysis of other respiratory infectious pathologies such as seasonal influenza outbreaks or future pandemic events.

In conclusion, this paper shows that with a simple multi-stage cleaning and processing pipeline, model performance can be boosted. We showed that categorizing CXR images systematically to avoid specific constraints associated with image acquisition can avoid biases in model prediction and result in boosted performance. We also showed that the boosted performance is based on pathological signals within the lung thereby emphasising the trustworthiness of the proposed pipeline from a clinical perspective.

**Acknowledgements** This research was supported by Wellcome Trust Clinical Research Career Development Fellowship 209553/Z/17/Z and the NIHR UCLH Biomedical Research Centre, UK. For the purpose of open access, the author has applied a CC-BY public copyright licence to any author accepted manuscript version arising from this submission.

## References

1. Apostolopoulos, I.D., Mpesiana, T.A.: Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and engineering sciences in medicine* **43**(2), 635–640 (2020)
2. Cushnan, D., Bennett, O., Berka, R., Bertolli, O., Chopra, A., Dorgham, S., Favaro, A., Ganepola, T., Halling-Brown, M., Imreh, G., et al.: An overview of the national covid-19 chest imaging database: data quality and cohort analysis. *GigaScience* **10**(11), giab076 (2021)
3. Cushnan, D., Berka, R., Bertolli, O., Williams, P., Schofield, D., Joshi, I., Favaro, A., Halling-Brown, M., Imreh, G., Jefferson, E., et al.: Towards nationally curated data archives for clinical radiology image analysis at scale: Learnings from national data collection in response to a pandemic. *Digital Health* **7**, 20552076211048654 (2021)
4. DeGrave, A.J., Janizek, J.D., Lee, S.I.: Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence* **3**(7), 610–619 (2021)
5. 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. IEEE (2009)
6. Gunraj, H., Sabri, A., Koff, D., Wong, A.: Covid-net ct-2: Enhanced deep neural networks for detection of covid-19 from chest ct images through bigger, more diverse learning. arXiv preprint [arXiv:2101.07433](https://arxiv.org/abs/2101.07433) (2021)
7. Heidari, M., Mirniaharikandehei, S., Khuzani, A.Z., Danala, G., Qiu, Y., Zheng, B.: Improving the performance of cnn to predict the likelihood of covid-19 using chest x-ray images with preprocessing algorithms. *International journal of medical informatics* **144**, 104284 (2020)
8. Jacob, J., Alexander, D., Baillie, J.K., Berka, R., Bertolli, O., Blackwood, J., Buchan, I., Bloomfield, C., Cushnan, D., Docherty, A., et al.: Using imaging to combat a pandemic: rationale for

- developing the UK national covid-19 chest imaging database. *European Respiratory Journal* **56**(2) (2020)
9. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D.: Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine* **17**(1), 1–9 (2019)
  10. Khan, A.I., Shah, J.L., Bhat, M.M.: Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images. *Computer Methods and Programs in Biomedicine* **196**, 105581 (2020)
  11. Organization., W.H.: Coronavirus disease (covid-19): Variants of sars-cov-2 (2021)
  12. Ozturk, T., Talo, M., Yildirim, E.A., Baloglu, U.B., Yildirim, O., Acharya, U.R.: Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in biology and medicine* **121**, 103792 (2020)
  13. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems* 32, pp. 8024–8035. Curran Associates, Inc. (2019), <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
  14. Rahimzadeh, M., Attar, A.: A modified deep convolutional neural network for detecting covid-19 and pneumonia from chest x-ray images based on the concatenation of xception and resnet50v2. *Informatics in medicine unlocked* **19**, 100360 (2020)
  15. Selvan, R., Dam, E.B., Detlefsen, N.S., Rischel, S., Sheng, K., Nielsen, M., Pai, A.: Lung segmentation from chest x-rays using variational data imputation. *arXiv preprint arXiv:2005.10052* (2020)
  16. 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: *Proceedings of the IEEE international conference on computer vision*. pp. 618–626 (2017)
  17. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: *International conference on machine learning*. pp. 6105–6114. PMLR (2019)
  18. Toraman, S., Alakus, T.B., Turkoglu, I.: Convolutional capsnet: A novel artificial neural network approach to detect covid-19 disease from x-ray images using capsule networks. *Chaos, Solitons & Fractals* **140**, 110122 (2020)

# 3D-3D Rigid Registration: A Comparative Analysis Study on Femoral Bone Scans



Perrine Solt, Adlane Habed, Antoine Bautin, Pierre Maillet,  
and Michel de Mathelin

**Abstract** 3D point cloud registration is an important step in a variety of computer-assisted surgeries and particularly critical for their success. Such registration is traditionally carried out using the Iterative Closest Point (ICP) algorithm although more recent and promising algorithms have been reported in the literature. In this paper, we provide a comparative analysis of several rigid registration algorithms for 3D point clouds including point-to-point ICP, point-to-plane ICP, Go-ICP and Super4PCS, in the difficult context of bone registration. In particular, we study the case in which a point cloud of femoral condyles is to be registered at the distal extremity of a human femoral bone. The condyles can typically be acquired during surgery using 3D sensors (e.g. intraoperative CT-scans, time-of-flight cameras, structured-light scanners) while the bone can be scanned preoperatively. The algorithms have been tested for speed, accuracy and robustness using both real and simulated data.

**Keywords** Point cloud registration · Femoral bone · ICP · Go-ICP · Super4PCS

---

P. Solt (✉) · A. Habed · M. de Mathelin  
CNRS, University of Strasbourg, Paris, France  
e-mail: [Perrine.solt@etu.unistra.fr](mailto:Perrine.solt@etu.unistra.fr)

A. Habed  
e-mail: [habed@unistra.fr](mailto:habed@unistra.fr)

M. de Mathelin  
e-mail: [demathelin@unistra.fr](mailto:demathelin@unistra.fr)

P. Solt · A. Bautin · P. Maillet  
Zimmer Biomet Robotics, Montpellier, France  
e-mail: [antoine.bautin1@zimmerbiomet.com](mailto:antoine.bautin1@zimmerbiomet.com)

P. Maillet  
e-mail: [pierre.maillet@zimmerbiomet.com](mailto:pierre.maillet@zimmerbiomet.com)

## 1 Introduction

In computer-assisted surgery, imaging modalities (e.g. CT-scans, MRI, X-Ray) and computer technologies provide the surgeon with a 3D representation of the surgical region of interest. This not only allows for a precise navigation during surgery, but also for the preoperative planning of the surgical procedure. However, for planning to be any useful intraoperatively, proper registration of the patient and the preoperative surgical data in a single reference frame is required. In essence, such registration consists in aligning a partial and noisy intraoperative 3D source point cloud on the complete and precise preoperative 3D model point cloud.

When the navigation system or the surgical robotic device is rigidly attached to the patient, registration may only be needed once, as an initial step. This is generally achieved using bone fiducials or inserts [6]. In this case the registration time, though important to reduce the overall duration of surgery, is not critical. However, when the patient moves during surgery, registration is to be renewed, in real time, of the order of a few tens of milliseconds, as to keep the patient and the planning (hence the preoperative 3D model) reference frames aligned at all times.

Most available commercial robotic and navigation systems use optical trackers that are rigidly fixed (e.g. with pins or screws) to the patient's bones so as to track their movements using an optical device [3, 19, 20, 22, 23]. This makes the tracking problem quite straightforward. Unfortunately, attaching the trackers to the bones is quite invasive and may lead to infections and/or fractures in addition to longer surgical procedures and healing time. RGB-D cameras offer a promising noninvasive alternative to the existing optical tracking systems. Indeed, such cameras are inexpensive and provide a relatively accurate 3D shape that may facilitate tracking and avoid using markers altogether. As a result, several works in progress investigate the use of RGB-D cameras for tracking [7–10, 17] in order to overcome the drawbacks of existing systems. Tracking then boils down to registering the 3D intraoperatively scanned bone portions on a preoperative 3D bone model. However, this requires the registration process to be fast as to allow real-time tracking. Furthermore, registration accuracy, robustness to noise as well as to different amounts of bone motion are also required for a safe and precise surgical act. The context of femoral bone surgery, which is the subject of the current study, is particularly difficult and challenging for such registration because only a small portion of the bone is visible during surgery [7].

This paper provides a comparative analysis study of 3D-3D rigid registration algorithms of bones in the context of knee surgery. In particular, we focus on the application described by Liu and Baena in [10] and in which femoral condyles are scanned with a RGB-D camera and require intraoperative registration on a preoperatively scanned femoral bone point cloud. This is a challenging registration procedure with real-time requirement and in which only a portion of the bone is visible during surgery. In this regard, we compare the viability and efficiency of the well-known and widely used Iterative Closest Point Algorithm (ICP) in its point-to-point [2] and point-to-plane [11] variants against two more recent and promising algorithms:

Super 4-Points Congruent Sets (Super4PCS) [12] and Globally Optimized ICP (Go-ICP) [21]. Super4PCS is based on the 4PCS procedure: it computes the best alignment in the least squares sense by finding coplanar bases of 4 congruent pointsets within a RANSAC (Random Sample Consensus) procedure. In contrast, Go-ICP is based on a branch-and-bound search using ICP as a subroutine. To test these algorithms in the context of the chosen application, we propose a workflow to simulate bone movement and evaluate the registration accuracy, while applying increasing perturbations to the scanned femoral condyles.

Our paper is organized as follows: the targeted registration algorithms are presented in Sect. 2. Section 3 describes the material and methods used to compare the registration algorithms. The results of our experiments along with our analysis are given in Sect. 4. Section 5 concludes this work and provides future works.

## 2 Registration Algorithms

This section provides a review of the four 3D-3D rigid point cloud registration algorithms considered in our comparative analysis study: namely, point-to-point ICP, point-to-plane ICP, Super4PCS and Go-ICP.

**Point-to-Point ICP:** The ICP algorithm from Besl and McKay [2] is the most commonly used algorithm for solving the 3D-3D registration problem and has many variants [16]. The algorithm attempts to register the two point clouds by iteratively considering the closest points (in the Euclidean sense) in the source and model point clouds as corresponding ones. For  $S$  a source point cloud containing  $N$  points with  $\mathbf{s}_i = (s_{ix}, s_{iy}, s_{iz}, 1)^T$  a source point, and  $\mathbf{m}_i = (m_{ix}, m_{iy}, m_{iz}, 1)^T$  the closest model point, at each iteration, the ICP algorithm estimates the optimal  $4 \times 4$  rigid transformation matrix  $\mathbf{M}_{opt}$  by solving

$$\mathbf{M}_{opt} = \arg \min_M \sum_{i=0}^{N-1} \|\mathbf{m}_i - \mathbf{M}\mathbf{s}_i\|^2. \quad (1)$$

The estimated transformation is then applied to the source points and the process of solving (1) is repeated.

**Point-to-Plane ICP:** While the point-to-plane variant of the ICP proceeds in the same iterative way as its point-to-point counterpart, it differs from it in that it considers, as a minimization metric, the distance between each source point and the tangent plane at its closest corresponding model point [11]. This allows one to take advantage of the surface normal information and was observed to generally lead to more robust and accurate registration results. Using the same notation as for the point-to-point ICP, and with  $\mathbf{n}_i = (n_{ix}, n_{iy}, n_{iz}, 0)^T$  the unit normal vector at point  $\mathbf{m}_i$  from the model, the optimal transformation is obtained by solving:



$$\mathbf{M}_{opt} = \arg \min_M \sum_{i=0}^{N-1} ((\mathbf{m}_i - \mathbf{M}\mathbf{s}_i)^T \mathbf{n}_i)^2. \quad (2)$$

**Super4PCS:** The 4PCS procedure consists in three steps [1]: creating some wide 4-points coplanar base, searching for all congruent bases and finding the most appropriate one. First, to create a coplanar base, three points, say  $p_1$ ,  $p_2$  and  $p_3$ , are randomly selected and a fourth point,  $p_4$ , is selected on the plane defined by the first three points. The size of this wide base is conditioned by the *overlap* value, set by the user. This value defines the proportion of common points in the point clouds. Then, congruent base points are extracted. For a rigid alignment, two distances are computed from the base as invariants. As it is always possible to find two intersecting lines between the four coplanar points, let set  $p_1p_2$  and  $p_3p_4$  the lines intersecting in a point  $p_5$ . The two invariants are defined by the ratios:

$$r_1 = \| p_1 - p_5 \| / \| p_1 - p_2 \| \quad (3a)$$

$$r_2 = \| p_3 - p_5 \| / \| p_3 - p_4 \| \quad (3b)$$

All the bases having the same invariants, up to a user-defined approximation level  $\delta$ , are selected. Finally, the best aligning transformation is sought within a RANSAC procedure. The chosen base is the one having the largest number of points within  $\delta$  distance from model points. In order to deal with the quadratic time complexity, Super4PCS removes the redundant 4-points candidates by using a rasterization approach.

**Go-ICP:** Go-ICP uses a nested branch-and-bound (BnB) structure together with the point-to-point ICP minimization problem (1). The BnB structure consists in splitting the search intervals using a tree structure, and evaluating candidate solutions by comparison with lower and upper estimated bounds. In this case, the outer BnB loop explores the rotation space, whereas the inner one explores the translational component of the rigid transformation. The algorithm is based on defining a progressively tight underestimator of the globally optimal registration error within a parameter space interval. While this corresponds to the most optimistic registration cost, the most pessimistic one is provided by the traditional local point-to-point ICP. Clearly, when an optimistic cost is worse than the pessimistic one, its corresponding parameter interval may be safely dropped. This makes the algorithm globally convergent and guarantees global optimality (up to a predefined optimality threshold).

Three main parameters of Go-ICP can be set by the user: the MSE threshold defining the convergence threshold based on the mean of squared errors, the trimming factor used to manage outliers, and the size of the distance transform used to compute the closest distances for bound evaluation.

### 3 Material and Methods

#### 3.1 Material and Preprocessing

For this study, we use one right femur 3D model extracted from the database provided by Nolte et al. [14]. A preprocessing of this 3D bone model is applied. It consists in placing the bone in a well-defined femoral coordinate system, in defining a region of interest (ROI) and in upsampling the point cloud.

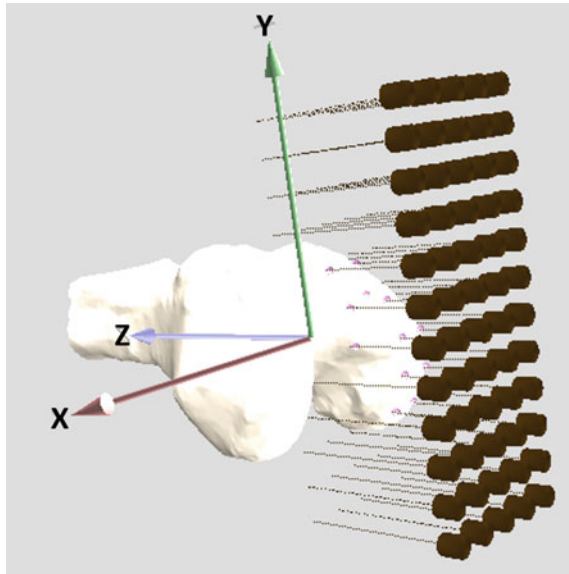
First, some studies propose methods to determine the anatomical axis of the femur and the tibia [4, 13]. For our purpose, we use the following definition (see Fig. 1):

- The X axis is defined as the epicondylar axis, going through both epicondylar points and oriented towards the right side of the patient, i.e. laterally for the right femur and medially for the left femur.
- The Y axis is placed on the antero-posterior axis, also named Whiteside line. It is oriented from the posterior to the anterior side.
- The Z axis is oriented from the intersection between X and Y axis towards the center of the femoral head.

Then, in order to decrease the computation time, a ROI is defined by keeping only the 10 cm femoral distal part of the point cloud. This is realistic since the part of the bone that is scanned during surgery is on the incision site.

Finally, point clouds are upsampled to increase and uniformize the point cloud density of the model. This is necessary because the fitness score computed by the

**Fig. 1** Femoral coordinate system and femoral condyles point cloud generation process. Only a half matrix of transducers with low density ( $d = 10$  mm) is presented for visualization ease



ICP algorithms depends on the point cloud density. For this purpose we use the Poisson-disk sampling algorithm presented by Corsini et al. [5] and implemented in Meshlab to create a Poisson density with a mean distance of 0.4 mm. After this preprocessing, the model point cloud contains 63,641 points.

### 3.2 Simulation Workflow

We propose an implementation of a simulation workflow. It aims at quantifying the registration error of the point cloud of the condyles on the preprocessed bone model point cloud during bone movement in real time. It consists of 4 steps: leg movement simulation, femoral condyles point cloud generation, registration and error quantification (see Fig. 2).

**Leg Movement Simulation:** The movement of the leg (Fig. 2a) is simulated through random transformations of the preprocessed bone model. These random transformations are generated with a Gaussian distribution of  $\mu_r$  mean Euler angle with  $\sigma_r$  standard deviation, and  $\mu_t$  mean translation with  $\sigma_t$  standard deviation. We denote by  $M_{applied}$  the resulting  $4 \times 4$  rigid transformation matrix. Different leg movement speeds are simulated by varying the mean value of these transformations. Note that, at each trial, the algorithms are fed the same data set obtained by applying a generated transformation.

**Femoral Condyles Point cloud Generation:** The point cloud of the femoral condyles is modeled by the following method (see Fig. 1). Each sensor is represented by a matrix of  $n \times m$  transducers. A raycast is generated from each transducer. All the intersections between the raycasts and the bone model mesh are collected and account for the condyle point cloud. Each condyle point belongs to the bone mesh, but is not necessarily included in the bone point set.

We propose to simulate different point cloud densities to account for various resolutions. This is done by varying the distance  $d$  (in mm) between two transducers (e.g. between two rays). We also add perturbation to this point cloud through Gaussian noise of mean value  $\mu_{noise}$  and standard deviation  $\sigma_{noise}$ . After this step, the source

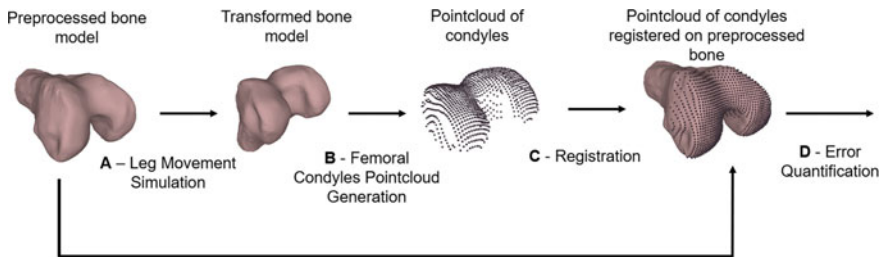


Fig. 2 Simulation workflow to compare registration error for different registration algorithms

point cloud is composed of 691 points for  $d = 2$  mm, 110 points for  $d = 5$  mm and 2747 points for  $d = 1$  mm.

**Registration:** The generated femoral condyle point cloud is registered (Fig. 2c) on the bone model point cloud using one of the registration algorithms described in Sect. 2.

**Error Quantification:** In order to quantify the registration error (Fig. 2d), we use a RMSE with the following definition.

For  $S$  a source point cloud containing  $N$  points, with  $s_i = (s_{ix}, s_{iy}, s_{iz}, 1)^T$  a source point and  $M_{est}$  the  $4 \times 4$  rigid transformation matrix estimated by the registration algorithm as the inverse of  $M_{applied}$  to align the transformed source on the model, the RMSE is obtained by:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} \|s_i - M_{est} M_{applied} s_i\|^2} \quad (4)$$

This value differs from the standard RMSE error internally used by the registration algorithms and also named fitness score [2]. In fact, the fitness score compares the distance from each point of the registered source to its nearest neighbor in the model. The RMSE instead compares the position of each point of the registered source (after a transformation has been applied)  $M_{est} M_{applied} s_i$  to the same point  $s_i$  from the initially well-aligned source.

This metric has two advantages: it is not influenced by the quality of the matching between both model and registered source point clouds and it does not depend on the model point cloud density. But the metric can only be used if the initially applied transformation is known, which is the case in our experiments.

### 3.3 Preprocessing

A point cloud preprocessing is applied before registration. First the model and the source point clouds are centered. For both point-to-point and point-to plane ICP, the parameters are computed so that the model is centered at the origin, and the same parameters are applied to the source point cloud. For Super4PCS and Go-ICP, we use an independent centralization: both source and model point clouds are centered at their respective origins. In all cases, a normalization is then applied such that the model lies within the unit-radius sphere. Both model and source point clouds are scaled with the same scale factor. Finally, the point indices in the model and source point clouds are randomized.

## 4 Experiments and Results

The simulations were carried out on a laptop with an Intel i7-11850H 2.50GHz CPU, Ubuntu 20.04, Oracle VM VirtualBox.

### 4.1 Parametrization of the Registration Algorithms

**Point-to-Point and Point-to-Plane ICP:** Point-to-Point and Point-to-plane ICP were tested using their respective Point Cloud Library's (PCL v1.12 [18]) implementations. These are based on solving (1) and (2) using Singular Value Decomposition.

**Parametrization of Super4PCS:** Super4PCS was tested using its OpenGR library implementation [15]. Particular attention was paid to the setting of following parameters:

- The *overlap* parameter defines the overlap ratio between the source and the model point clouds, with respect to surface area of the smallest point cloud. The number of trials of different 4-point sets bases is directly linked to this value: the larger *overlap*, the less trials. The femoral condyles being totally included in the bone point cloud, the real *overlap* value is 1. In our experiments, the *overlap* was set to 0.9 because in practice some points may not necessarily belong to the bone. As a consequence, the number of RANSAC iterations naturally increases.
- The parameter  $\delta$  has an influence on the process of extraction of the pairs of points and in the search of congruent bases. With a larger  $\delta$ , more pairs of points are considered to have the same invariants and more congruent bases are evaluated. For  $\delta$  too small, the number of retrieved potential correspondences may be insufficient and the algorithm fails to converge. With a much bigger  $\delta$  than it ought to be, many more possibilities are evaluated and the algorithm may take prohibitively longer to terminate. In our experiments  $\delta$  was set to 1 mm and then normalized with the same scale factor used during the normalization process described in Sect. 3.3.
- Super4PCS relies on a sparse matching of points across point clouds. It is hence essential, as it is also recommended in the OpenGR library documentation, to use a reasonable sample size limited to a few thousands of points. We used a sample size of 3000 points.
- We set 60 s as a time limit as it has been proposed for some test data provided by Mellado et al. [12] along with their software.

As suggested by the authors, the registration output of Super4PCS is fed into an ICP refinement.

**Parametrization of Go-ICP:** Some parameters require proper setting by the user for the Go-ICP algorithm.

- The *trimming* factor is meant to manage outliers by excluding extreme values. We set it to 0.

- Go-ICP's runtime depends on the convergence threshold (global optimality gap). A compromise is hence needed between time and accuracy. We set the optimality threshold to 0.001 mm. A smaller value slows the algorithm down but increases the registration accuracy.
- The number of nodes per dimension of the distance transform is used to compute the closest distances for fast bound evaluation. We set this value to 30.

**Description of the experiments:** We conducted four experiments. For each of them, we were interested in the mean registration time (Fig. 3), in the convergence (Fig. 4) and in the mean of the RMSE value (Fig. 6) over 100 trials.

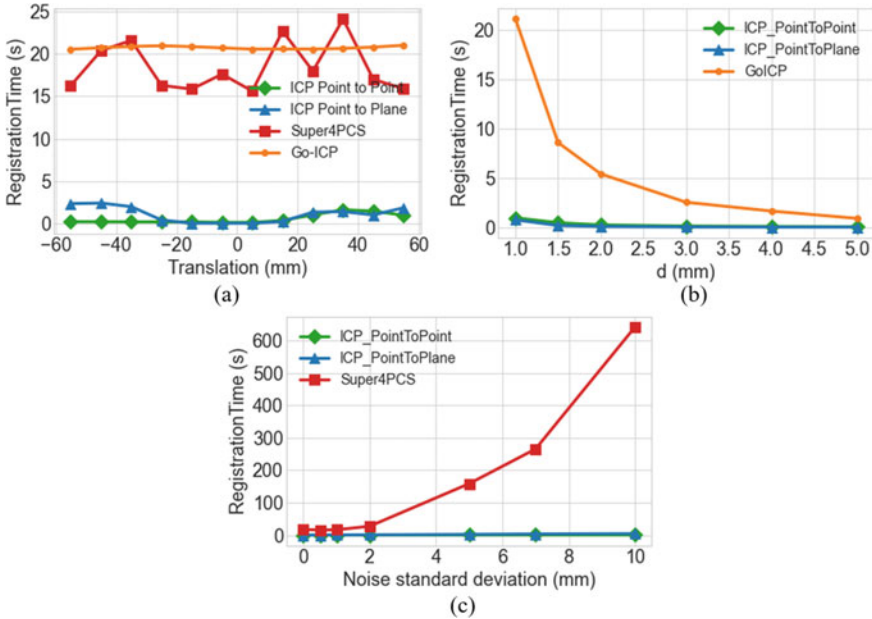
Both point clouds were initially aligned. For the two first experiments, we were interested in the ability of the algorithms to converge while increasing the bone movement amplitude for a fixed time difference. This movement is defined as a transformation composed of a rotation and a translation. Except for the third experiment, the matrix of transducers used to generate the femoral condyle points is of size of  $100 \times 120$  boxes, with  $d = 2$  mm. All the distances, for applied translations and distance between transducers, are defined in mm before being normalized with the same scale factor used during the normalization process described in Sect. 3.3.

- First, we increased the translation while not applying any rotation.  $\mu_t$  was taken in the range between  $-55$  and  $55$  mm with a step of  $10$  mm along the three axes X, Y and Z, with a fixed  $\delta_t = 5$  mm.
- We then assessed the robustness of the algorithms while increasing the rotation with a small realistic Gaussian translation ( $\mu_t = 5$  mm,  $\sigma_t = 5$  mm).
- In the third experiment, we focused on the influence of the dimensions of the matrix of transducers on the registration convergence. We increased  $d$  to change the density of the source point cloud. Gaussian rotations ( $\mu_r = 10^\circ$ ,  $\sigma_r = 5^\circ$ ) and translations ( $\mu_t = 10$  mm,  $\sigma_t = 5$  mm) were applied.
- Finally, we assessed the impact of noise on each registration algorithm. For this purpose, we increased  $\sigma_{noise}$  while keeping  $\mu_{noise} = 0$  mm. Gaussian rotations ( $\mu_r = 10^\circ$ ,  $\sigma_r = 5^\circ$ ) and translations ( $\mu_t = 10$  mm,  $\sigma_t = 5$  mm) were applied.

## 4.2 Registration Time

Figure 3a shows that Go-ICP and Super4PCS were quite slow (between 15 and 25 s) in comparison to the local algorithms point-to-point and point-to-plane ICP (less than 4 s) in case of variation of translations. For translation values between  $-15$  mm and  $15$  mm (resp.  $-5$  mm and  $5$  mm), point-to-point ICP converged on average in 0.26 s (resp. 0.18 s), and point-to-plane ICP converged on average in 0.13 s (resp. 0.08 s).

It can be seen in Fig. 3b that, by reducing the distance  $d$  between the center of each transducer (i.e. by increasing the source point cloud density), the registration time increased exponentially for Go-ICP, while it did not increase much for both local ICP variants.



**Fig. 3** Registration time for: **a** a variation of the translation applied to the source point cloud, **b** a variation of the distance between each transducer, **c** a variation in the applied noise standard deviation

While applying increasing noise to the source point cloud (see Fig. 3c), the registration time did not increased much for point-to-point ICP: typically 0.24 s of mean time in the absence of noise and 0.58 s in the presence of noise with a standard deviation of 10 mm. Point-to-plane ICP registration time increased from 0.17 s with no noise to 4.61 s for noise with 10 mm standard deviation. Super4PCS was a lot slower with a mean registration time of 17.2 s without noise and 642 s for noise with 10 mm standard deviation. The simulations with Go-ICP were not carried out to the end because, for each trial, the time exceeded 95 s for a noise standard deviation of 2 mm and to more than 2 h for noise with 5 mm standard deviation.

### 4.3 Convergence

We are interested in the convergence of the different algorithms. We consider that a run converges if the RMSE value is less than 5 mm. This value has been chosen by considering the histograms from Fig. 4, with a bin size of 5 mm. It separates the cluster of points with a RMSE close to 0 from the rest of the clusters of points with high RMSE values, representing point clouds that have been wrongly registered.

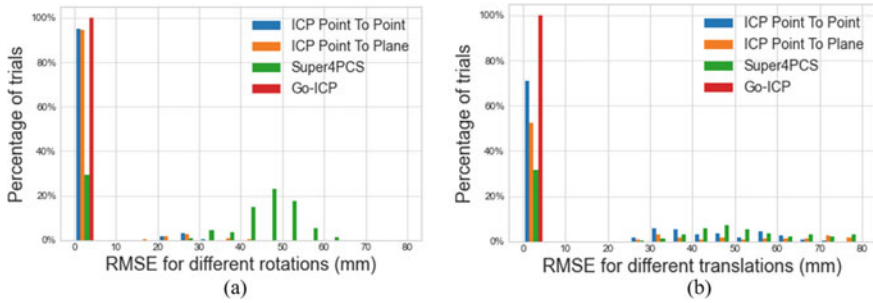


Fig. 4 Histogram of RMSE for: **a** a variation in the rotation, **b** a variation in the translation

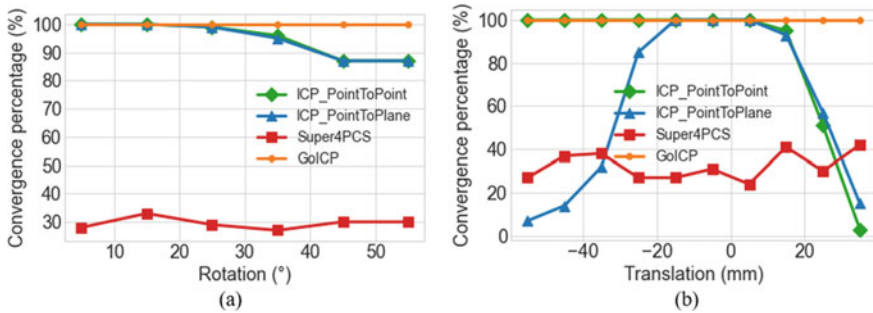


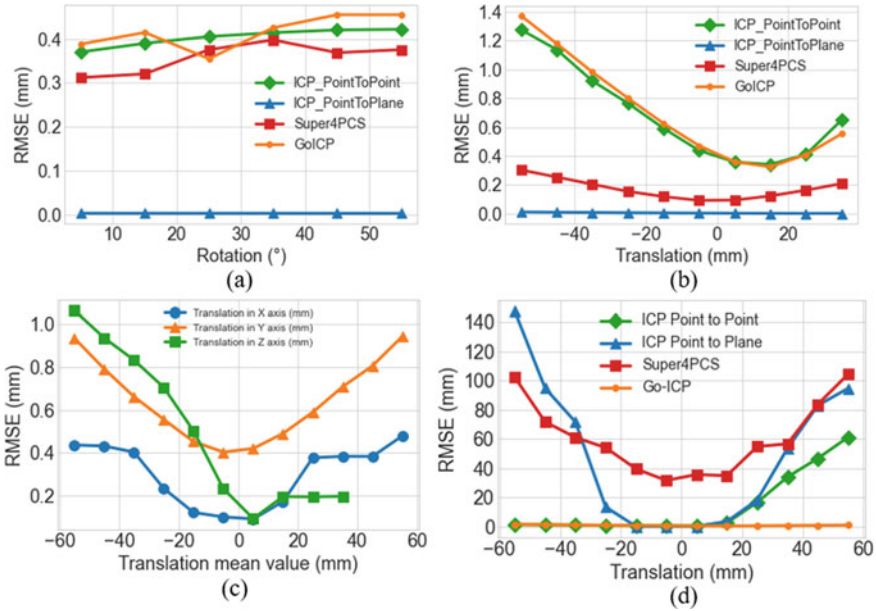
Fig. 5 Convergence percentage for: **a** a variation in the rotation, **b** a variation in the translation

In both cases of rotation variation (Fig. 5a) and translation variation (Fig. 5b), Super4PCS converged only in about 30% of the cases with our parametrization. Go-ICP always converged. Both local ICP algorithms increasingly failed to converge for rotations of more than 20° and for translations exceeding than 10 mm. Such failures are likely due to a premature convergence to a local optimum.

### 4.4 Registration Accuracy

Considering only the cases where the algorithms converged, Super4PCS exhibited the same behavior as that of Go-ICP and point-to-plane ICP when increasing the rotations (see Fig. 6a), with a RMSE error above 0.3 mm, while point-to-plane ICP outperformed all algorithms with a RMS error of 3.3E−3 mm. For translations (see Fig. 6b), Super4PCS performed better than point-to-point ICP and Go-ICP when it converged. No values for translations more than 35 mm are shown because point-to-point ICP failed to converge all the time. To explain the shift in the convergence observed in Fig. 6b, we present in Fig. 6c the variation of the RMSE for the point-to-point ICP algorithm by decomposing the applied translations into each Euclidean axis. We suspect the shape of the bone point cloud to be the cause of the behavior observed on the Z axis, and to induce the previously mentioned shift. Figure 6d shows the mean RMS error for all the algorithms with non-converging cases included.





**Fig. 6** Registration accuracy for: **a** a variation of the only rotation in the case of convergence, **b** a variation in only the translations in the case of convergence, **c** a variation of the translation in Euclidean axes only in case of convergence, **d** a variation in the translations for all cases (i.e. with and without convergence)

## 5 Conclusion and Future Work

We conducted a comparative analysis study of four 3D-3D rigid registration algorithms—point-to-point ICP, point-to-plane ICP, Super4PCS and Go-ICP—in the context of knee tracking with a 3D camera. We tested the registration robustness to the amplitude of the leg movement (for increasing transformations), to noise and to the density of the source point cloud. Our study has shown that point-to-plane ICP is the most adequate to guarantee convergence despite a variation in the source point cloud density and noise, while being fast. The condition for that is that the movement amplitude stays small, thus inducing small motions between consecutive point clouds. This is realistic because of the high frequency of data acquisition in real-time tracking. Nevertheless, our study shows that there is a need for registration algorithms that are more suitable for real-time applications. Dedicated solutions, such as local point cloud descriptors, specific to the bone local shape, should be investigated.

## References

1. Aiger, D., Mitra, N., Cohen-Or, D.: 4-points congruent sets for robust pairwise surface registration. 35th International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'08) 27 (08 2008)
2. Besl, P., McKay, H.: A method for registration of 3-D shapes. Pattern Analysis and Machine Intelligence, IEEE Transactions on 14, 239–256 (1992)
3. Brainlab: Knee3 surgical technique p. 58 (2015), <https://www.brainlab.com/wp-content/uploads/2016/12/Knee3-Surgical-Technique.pdf>
4. Chen, H., Kluijtmans, L., Bakker, M., Dunning, H., Kang, Y., van de Groes, S., Sprengers, A.M.J., Verdonschot, N.: A robust and semi-automatic quantitative measurement of patellofemoral instability based on four dimensional computed tomography. Medical Engineering & Physics 78, 29–38 (2020)
5. Corsini, M., Cignoni, P., Scopigno, R.: Efficient and Flexible Sampling with Blue Noise Properties of Triangular Meshes. IEEE transactions on visualization and computer graphics 18, 914–24 (2012)
6. Fitzpatrick, J.M.: The Role of Registration in Accurate Surgical Guidance. Proceedings of the Institution of Mechanical Engineers. Part H, Journal of engineering in medicine 224(5), 607–622 (2010)
7. He, G., Mustahsan, V.M., Bielski, M.R., Kao, I., Khan, F.A.: Report on a novel bone registration method: A rapid, accurate, and radiation-free technique for computer- and robotic-assisted orthopedic surgeries. Journal of Orthopaedics 23, 227–232 (2021)
8. Hu, X., Liu, H., Rodriguez y Baena, F.: Markerless Navigation System for Orthopaedic Knee Surgery: A Proof of Concept Study. IEEE Access 9 (Apr 2021)
9. Hu, X., Nguyen, A., Baena, F.R.y.: Occlusion-robust visual markerless bone tracking for computer-assisted orthopedic surgery. IEEE Transactions on Instrumentation and Measurement 71, 1–11 (2022)
10. Liu, H., Baena, F.R.Y.: Automatic Markerless Registration and Tracking of the Bone for Computer-Assisted Orthopaedic Surgery. IEEE Access 8, 42010–42020 (2020)
11. Low, K.L.: Linear least-squares optimization for point-to-plane ICP surface registration (01 2004), [https://www.comp.nus.edu.sg/~lowkl/publications/lowk\\_point-to-plane\\_icp\\_techrep.pdf](https://www.comp.nus.edu.sg/~lowkl/publications/lowk_point-to-plane_icp_techrep.pdf)
12. Mellado, N., Aiger, D., Mitra, N.J.: Super 4PCS Fast Global Pointcloud Registration via Smart Indexing. Computer Graphics Forum 33(5), 205–215 (2014)
13. Miranda, D.L., Rainbow, M.J., Leventhal, E.L., Crisco, J.J., Fleming, B.C.: Automatic Determination of Anatomical Coordinate Systems for Three-Dimensional Bone Models of the Isolated Human Knee. Journal of biomechanics 43(8), 1623–1626 (2010)
14. Nolte, D., Tsang, C.K., Zhang, K.Y., Ding, Z., Kedgley, A.E., Bull, A.M.: Non-linear scaling of a musculoskeletal model of the lower limb using statistical shape models. Journal of Biomechanics 49(14), 3576–3581 (2016)
15. OpenGR (Aug 2022), <https://github.com/STORM-IRIT/OpenGR>, original-date: 2018-04-23T06:36:52Z
16. Pomerleau, F., Colas, F., Siegwart, R., Magnenat, S.: Comparing ICP variants on real-world data sets Open-source library and experimental protocol. Autonomous Robots 34(3), 133–148 (2013), publisher: Springer
17. Rodrigues, P., Antunes, M., Raposo, C., Marques, P., Fonseca, F., Barreto, J.P.: Deep segmentation leverages geometric pose estimation in computer-aided total knee arthroplasty. Healthcare Technology Letters 6(6), 226–230 (2019)
18. Rusu, R.B., Cousins, S.: 3D is here: Point Cloud Library (PCL). In: IEEE International Conference on Robotics and Automation (ICRA). Shanghai, China (May 9–13 2011)
19. Smith&Nephew: Navio surgical system, surgical technique for total knee arthroplasty p. 44 (2018), <https://www.smith-nephew.com/global/surgicaltechniques/recon/14529%20v1%20500095%20revc%20navio%20ka%20surgical%20technique%200718.pdf>

20. Stryker: Mako@pka, application user guide p. 108 (2018), <https://www.strykermeded.com/media/2943/mako-pka-application-user-guide-pka-30-003.pdf>
21. Yang, J., Li, H., Campbell, D., Jia, Y.: Go-ICP: A Globally Optimal Solution to 3D ICP Point-Set Registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(11), 2241–2254 (Nov 2016), [arXiv:1605.03344](https://arxiv.org/abs/1605.03344) [cs]
22. ZimmerBiomet: Othosoft®, knee 2.4 universal p. 66, <https://www.zimmerbiomet.com/content/dam/zimmer-biomet/medical-professionals/knee/optical-navigation-system/orthosoft-knee-2-4-universal-user-manual.pdf>
23. ZimmerBiomet: Rosa@knee system, surgical technique v1.1 p. 60 (2020), [https://www.rosaknee.com.br/assets/pdfs/ROSA%20Knee%20Surgical%20Technique%20v1.1%20\(ingles\).pdf](https://www.rosaknee.com.br/assets/pdfs/ROSA%20Knee%20Surgical%20Technique%20v1.1%20(ingles).pdf)

# Fully Automatic Axial Vertebral Rotation Measurement of Children with Scoliosis Using Convolutional Neural Networks



Jason Wong , Marek Reformat , and Edmond Lou 

**Abstract** Adolescent idiopathic scoliosis is a three-dimensional spinal disorder, where the spine is characterized by lateral curvature and axial vertebral rotation (AVR). Measurement of AVR is not as common as the lateral curvature due to its time-consuming nature. However, AVR measurements are useful for predicting curve progression and planning surgeries, which could both result in improved treatment outcomes. To improve accessibility to AVR measurements, this study reported on a convolutional neural network-based method that automatically measured the AVR on posteroanterior (PA) radiographs based on Stokes' method. The proposed method was tested on 26 PA radiographs (338 vertebrae). The method resulted in 84% of automatic measurements within the clinically accepted error of  $5^\circ$  and achieved a circular mean absolute error of  $3.1^\circ \pm 3.5^\circ$  when compared with manual measurements. This high accuracy, coupled with quick computation time (1.7 s per vertebra) and highly interpretable outputs, demonstrates the clinical feasibility of employing the proposed automatic method. This is the first method that automatically measures AVR accurately on PA radiographs taken by both the conventional and EOS x-ray imaging systems.

**Keywords** Axial vertebral rotation · Convolutional neural network · Machine learning · Radiograph · Scoliosis

## 1 Introduction

Adolescent idiopathic scoliosis (AIS) is a three-dimensional (3D) spinal disorder, where the spine features a coronal curvature, abnormal sagittal curvature, and axial vertebral rotation (AVR). This disorder occurs in adolescents aged 10–16 years old and affects 1–3% of this population. The severity of AIS is typically quantified using only the Cobb angle, a measure of the lateral curvature of the spine [1]. However, only measuring the Cobb angle may underestimate the 3D spinal deformation and delay

---

J. Wong · M. Reformat · E. Lou (✉)  
University of Alberta, Edmonton, Alberta T6G 1H9, Canada  
e-mail: [elou@ualberta.ca](mailto:elou@ualberta.ca)

treatment management. AVR is another scoliotic parameter that has been found to be relevant for treatment and prognosis of AIS. Providing AVR measurements to a clinician would allow for a better understanding of the progression of a subject's AIS. This could allow for earlier treatment, maximizing treatment effectiveness and minimizing surgical intervention. Also, accurate AVR measurements is crucial for surgical planning. Incorrectly evaluating a subject's AVR could increase the chances of pedicle breaches during surgery, resulting in a higher risk of screw-related complications [2].

Ideally, AVR is measured using a computerized tomography (CT) scan because the full 3D spine is imaged and the AVR can be directly measured. However, due to its high ionizing radiation dosage, CT imaging is not used regularly. From literature, the most common method which can quantify the AVR on PA radiographs is Stokes' method. It consists of identifying the minimum width of the vertebral body along with the centroids of the pedicles. Actual 3D information is then incorporated into their formula to calculate AVR by using pre-computed vertebra width-depth ratios [3]. However, labelling the pedicles is particularly difficult and time-consuming because the area around them is sometimes unclear due to poor contrast on the radiograph.

Consequently, a computer-assisted method of AVR measurement that is fully or even partially automated is widely sought by clinicians [4]. Two other groups have reported on automating steps of AVR measurement on PA radiographs [5, 6]. However, their methods still required users to manually identify spinal features, such as the general area of the pedicles, spinal curve, or vertebral endplates. No other group has developed a fully automatic AVR measurement method for PA radiographs.

Our group has previously developed a machine learning algorithm for PA radiographs, which achieved a circular mean  $\pm$  standard deviation of absolute errors of  $4.3^\circ \pm 5.7^\circ$  [7]. Among the 221 tested vertebrae, 81% of automatic measurements were within the clinically accepted error of  $5^\circ$  when compared with manual measurements. However, that algorithm was developed to only work for radiographs taken by the EOS x-ray system. The EOS system (EOS Imaging, France) is a low-dose ionizing radiation x-ray system, which is commonly used at hospitals in developed countries. However, many private scoliosis centers and low-income countries are still employing the conventional x-ray system. Consequently, a machine learning algorithm that can measure the AVR from both the conventional and EOS x-ray systems is highly desired. This manuscript reported on a fully automatic algorithm that could measure the AVR on PA radiographs taken by both the conventional and EOS x-ray systems and on the accuracy of this automatic measurement method.

## 2 Methodology

### 2.1 Proposed Method

The proposed AVR measurement algorithm was accomplished using a cascaded convolutional neural network (CNN) design, where the algorithm was divided into six stages: (1) segmentation of the spinal column from the top thoracic (T1) to bottom lumbar (L5) vertebra, (2) segmentation of the individual vertebral bodies, (3) segmentation of the pedicle centroids, (4) iterative location of the vertebral body, (5) correction of the vertebral body segmentations, and (6) AVR measurement using Stokes' method. The flowchart of the procedure (Fig. 1) is modified from our previous work [7], by adding vertebral body segmentation correction to improve vertebral body width calculations.

Spinal column, vertebral body, and pedicle centroid segmentation were all accomplished with CNNs trained using the supercomputer from the Industry Sandbox and Artificial Intelligence Computing (ISAIC) at the University of Alberta. Training and validation were performed on a Linux virtual machine with an Intel Xeon Gold 6138 dual processor, NVIDIA Tesla V100 16 GB GPU, and 64 GB of RAM. The code for this study was developed in the Python language, using TensorFlow for CNN development.

To train CNNs and validate the final measurement algorithm, PA radiographs of children with AIS were extracted from local scoliosis clinical records. Ethics approval (Pro00102044) was granted by the University of Alberta research health ethics board.

#### 2.1.1 Stage 1—Segmentation of the Spinal Column

This segmentation stage crops out the key spinal column region to assist the next two stages of segmentation in locating the individual vertebral bodies and pedicles. This stage includes image processing, scaling the image down to  $256 \times 128$ , applying a variant of the U-net CNN, and post-processing to obtain the spinal column curve (SCC). The detailed description can be found in [7]. A new dataset and hyper-parameters were used in this study to train the spinal column segmentation CNN (CNN<sub>SC</sub>).

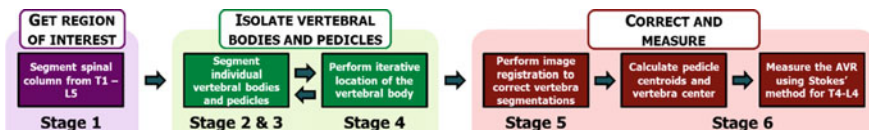
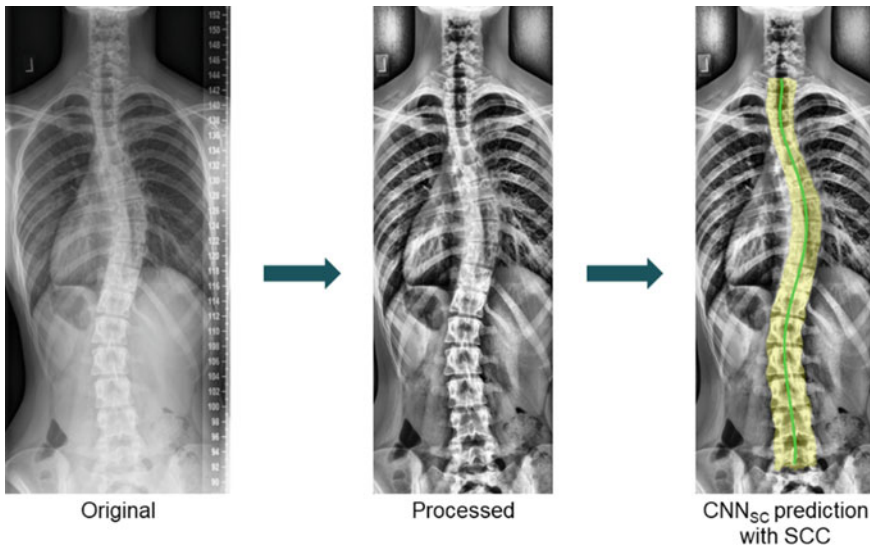


Fig. 1 Flowchart of the proposed automatic AVR measurement algorithm

To train and optimize  $CNN_{SC}$ , 110 PA radiographs were split into a 96-image training and 24-image validation set randomly. On all 110 images, the spinal column from T1 to L5 was labelled as one continuous segment. Both sets comprised of half conventional and half EOS radiographs. Data augmentation was employed to increase the effective size of the training set by randomly flipping the image horizontally or rotating the image by an angle of up to  $10^\circ$ .

A grid search was performed to optimize the hyperparameters of  $CNN_{SC}$ .  $CNN_{SC}$  aimed to minimize the soft Dice loss function and was trained using the Adam optimizer. The optimized  $CNN_{SC}$  was trained for 1,000 epochs with a  $10^{-3}$  learning rate and batch size of 2. The leaky ReLU activation function with 0.01 alpha was used after each convolutional layer [8]. The regularization techniques of batch normalization and dropout were performed to improve the network’s ability to generalize. Batch normalization was performed before each pooling and upsampling layer, and dropout after each pooling and upsampling layer with a probability of 0.5. The network with the lowest Dice loss on the validation set during training was treated as the final  $CNN_{SC}$ . Figure 2 depicts the image pipeline from initial PA radiograph to segmented spinal column with SCC.



**Fig. 2** PA radiograph pipeline from original to segmented spinal column (yellow) with the SCC (green)

### 2.1.2 Stage 2—Segmentation of the Vertebral Body

The vertebral body was segmented on square images that were cropped from the initial unprocessed PA radiographs and centered around the vertebral body. Segmentation of the vertebral body was required to obtain the center of the minimum width of the vertebral body ( $C_{VB}$ ). To highlight the boundaries of the vertebral body, the images underwent histogram equalization, and then were scaled down to a size of  $128 \times 128$ .

A similar CNN architecture ( $CNN_{VB}$ ) as  $CNN_{SC}$  was used to segment the vertebral body, with the only difference being an initial input image size of  $128 \times 128$ . A total of 20 PA radiographs were selected from the database, from which 340 vertebral body images were manually cropped, labelled, and randomly sampled to form a 272-image training and 68-image validation set. Each set consisted of half conventional and half EOS images. Random horizontal flipping, zooming from 80 to 120%, horizontal and vertical shifts of up to 10%, and/or rotations of up to  $45^\circ$  were employed as data augmentation methods so that  $CNN_{VB}$  was more robust to different vertebra anatomies.

The same grid search practices for  $CNN_{SC}$  were employed to optimize  $CNN_{VB}$ . Using the Adam optimizer and a soft Dice loss function,  $CNN_{VB}$  was trained for 1000 epochs with a  $10^{-4}$  learning rate and a batch size of 4. The leaky ReLU activation function with 0.01 alpha was used after each convolutional layer. Dropout was performed with 0.125 probability, and batch normalization was not performed. The network with the lowest Dice loss on the validation set was treated as the final  $CNN_{VB}$ .

### 2.1.3 Stage 3—Segmentation of the Pedicle Centroids

Using the same input images to  $CNN_{VB}$ , the pedicle centroids were identified using the same architecture ( $CNN_{PED}$ ) as  $CNN_{VB}$ . However,  $CNN_{PED}$  aimed to predict the centroids of the pedicles. Therefore, images were labelled such that the centroids of the two pedicles on the vertebral body image were digitized. Then, the labels were encoded for heatmap regression, where the labelled centroid pixel had a value of 1 and as you moved further from this center, the surrounding pixels decreased in value. How the values decreased was determined using a Gaussian neighborhood function with a standard deviation of 2 and a window of 3 (creating a  $7 \times 7$  square centered on each pedicle centroid). A total of 390 vertebral body images derived from 30 PA radiographs, with the pedicle centroids labelled, were randomly sampled to form a 312-image training and 78-image validation set. Each set consisted of half conventional and half EOS images. The same data augmentation practices as  $CNN_{VB}$  were employed for  $CNN_{PED}$ .

$CNN_{PED}$  was trained using a mean squared error loss function instead of soft Dice. Other than that, optimizing  $CNN_{PED}$  followed a similar grid search practice as the others. The optimized  $CNN_{PED}$  was trained using an Adam optimizer for 500 epochs with a learning rate of  $10^{-4}$  and a batch size of 1. The leaky ReLU activation function



with 0.01 alpha was used after each convolutional layer. Dropout of probability 0.125 was performed, and there was no batch normalization. The network with the lowest mean squared error loss on the validation set was treated as the final CNN<sub>PED</sub>.

#### 2.1.4 Stage 4—Iterative Location of the Vertebral Body

Because the inputs to CNN<sub>VB</sub> and CNN<sub>PED</sub> are square images centered on the vertebral body, a procedure to locate and crop out these images is required. An iterative algorithm was developed to localize these vertebrae for segmentation. The general structure of this iterative algorithm is like our previous work [7]: segmenting a starting vertebra at around T12, segmenting the vertebra directly above, evaluating the quality of the segmentation by using a metric derived from standard masks, cropping another image to re-segment if a certain quality was not achieved, and repeating this procedure iteratively up and down until the ends of the spinal column are reached. However, one major improvement was made to the iterative location algorithm.

The metrics with which to evaluate the quality of a vertebral body segmentation were modified. The previous mean and variance quality metrics were replaced with a single similarity loss coefficient ( $\ell$ ). Calculating  $\ell$  involved using thirteen standard vertebral body masks from the CNN<sub>VB</sub> training set to compare with a given predicted segmentation. First, the predicted segmentation was rotated such that its bounding box was flat with the horizontal. Then, its contour was extracted, along with the contours of each standard mask. The minimum distances from each point on the standard contour to the predicted contour and vice versa were then calculated. Let the set of all standard masks be  $\psi$  and the distribution of points in the standard and predicted contour be  $\Gamma$  and  $\Phi$  with a single point being  $\gamma$  and  $\varphi$  in each contour, respectively. The similarity loss  $\ell$  is calculated as follows:

$$\ell = \min_{\psi} \left[ \sqrt{\frac{1}{N_{\Gamma}} \sum_{j=1}^{N_{\Gamma}} \left( \min_{\varphi \in \Phi} \gamma_j - \varphi \right)^2} + \sqrt{\frac{1}{N_{\Phi}} \sum_{j=1}^{N_{\Phi}} \left( \min_{\gamma \in \Gamma} \varphi_j - \gamma \right)^2} \right] \quad (1)$$

where  $N_{\Gamma}$  and  $N_{\Phi}$  refer to the number of points in distributions  $\Gamma$  and  $\Phi$ , respectively. A higher value of  $\ell$  refers to a poorer vertebral body segmentation.

#### 2.1.5 Stage 5—Correction of the Vertebral Body Segmentation

A common trend in our previous algorithm's inaccurate measurements was poor vertebral body segmentation that influenced the position of  $C_{VB}$ . Cases where the segmentation contained an extra protrusion at the sides of the segmentation increased the chances of AVR measurements being outside of clinical acceptance due to the shift in the  $C_{VB}$  calculation. Therefore, an image registration step was incorporated to correct these segmentations and maximize the chances of obtaining an accurate  $C_{VB}$ .

The scaling iterative closest points algorithm (SICP) was implemented to accomplish image registration correction [9]. SICP is an algorithm that searched for the affine transformation (scaling, rotation, and translation) that minimized the sum of squared minimum distances between the points on two given contours. For our study, these two contours were of the predicted segmentation and the standard vertebral body mask that produced the lowest  $\ell$  in the iterative vertebral body location algorithm. SICP was repeated until an improvement of less than  $10^{-3}$  was obtained between consecutive iterations or until 100 iterations was reached. An inlier ratio of 0.8 was used. The registered standard mask was used in place of the predicted segmentation to calculate  $C_{VB}$ .

### 2.1.6 Stage 6—Measurement of the AVR

With the vertebral body segmented and the probability heatmap from  $CNN_{PED}$  obtained, the algorithm conducted post-processing to derive the necessary information for measuring AVR. Based on vertebral anatomy, the vertical position of the minimum width is located close to the centroid of the vertebral body. Therefore, to determine  $C_{VB}$ , the algorithm calculated the geometric centroid of the rotated mask and searched for the minimum width in a vertical window centered around the centroid's y-position. The algorithm then calculated the lengths of the largest continuous segment for each row within the window. The location of the median segment length was treated as the vertical position of the minimum width, and  $C_{VB}$  was then calculated by determining the x-centroid of this continuous segment.

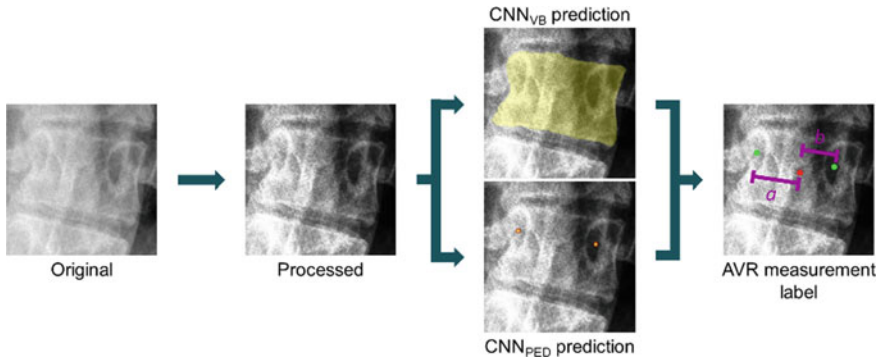
Next, the algorithm performed post-processing on the predicted probability heatmaps. First, a list of potential centroid candidates was determined through iterative local probability thresholding. This consisted of obtaining the connected components when the heatmap was thresholded by a value  $\theta$  and identifying the pixel with the highest probability within each connected component. These identified pixels were stored in a list of potential candidate centroids. This process was iterated for multiple  $\theta$ , starting at 0.5 and decrementing by 0.05 each time until 0.01 was reached.

With a list of potential centroid candidates obtained, the algorithm separated the candidates according to the different halves of the vertebral body (potential left and right pedicles), based on where they were relative to the vertebral body's centroid. If no left or no right candidates were found during thresholding, then the algorithm skipped the vertebra. Otherwise, the algorithm picked the candidate with the highest probability within each half. This procedure was repeated for all vertebrae from T4-L4 inclusive.

Using these points, the AVR is calculated as:

$$AVR = \arctan\left(\frac{w}{2d} \times \frac{a-b}{a+b}\right) \quad (2)$$

where  $w/d$  refers to the pre-computed width-depth ratios and  $a$  and  $b$  refer to the distances between the center of the minimum width and the centers of the pedicles.



**Fig. 3** Vertebral body image pipeline from original to AVR measurement labelled with pedicle centroids in green, and  $C_{VB}$  in red. In  $CNN_{VB}$ , yellow indicates the predicted segmented pixels and in  $CNN_{PED}$ , pixels from white to dark red indicate higher to lower probability predictions

Figure 3 depicts the image pipeline from original vertebral body to processed, CNN segmented, and AVR measurement labelled. The AVR was calculated for T4-L4 inclusive.

Finally, an additional post-processing step of verifying that the AVR values were reasonable was conducted. This involved first calculating confidence scores of each vertebral body based on the pedicle centroid probabilities and the  $\ell$  of the vertebral body mask. A maximum confidence score was obtained if both pedicle centroids were predicted with 100% probability and if the predicted vertebral body segmentation matched perfectly with one of the standard masks. Then, using the vertebral body with the highest confidence score as the baseline, the algorithm examined the vertebra above to check that its AVR values were reasonable. If there was a difference in AVR values between neighboring vertebrae of  $10^\circ$  or if the trend of AVR values did not follow a smooth curve, then a search for different pedicle centroid positions was conducted. This search consisted of calculating all possible AVR values from the previously identified left and right pedicle centroid candidates. The pair of centroid candidates that produced the AVR closest to what was expected based on the trend of the AVR values was chosen as the true centroids. With this vertebra now complete, the algorithm iterated upwards, treating the recently checked vertebra as the baseline and investigating the next vertebra above. This repeated until T4 was reached, and then the algorithm went back to the original baseline vertebra and repeated the same iterative procedure, but moving downwards instead until L4 was reached.

## 2.2 Validation

### 2.2.1 Spinal Feature Segmentation

To evaluate the performance of  $\text{CNN}_{\text{SC}}$  and  $\text{CNN}_{\text{VB}}$ , fivefold cross-validation was employed. This was conducted using the optimized hyperparameters found from the grid searches. Each fivefold cross-validation used all labelled images—110 spinal column images and 340 vertebral body images. Performance was evaluated using the means and standard deviations of precision, recall, and the Dice coefficient over the folds. Cross-validation was not performed for  $\text{CNN}_{\text{PED}}$  because the nature of its outputs is different and cannot be evaluated with intuitive performance metrics such as precision, recall, or the Dice coefficient.

### 2.2.2 AVR Measurement

To evaluate the performance of the automatic AVR measurement algorithm, a set of 26 spinal PA radiographs was selected. The set comprised of 13 conventional and 13 EOS radiographs, and all images were randomly selected. The average AVR of the 26 subjects for all valid vertebrae was  $3.4^\circ \pm 2.9^\circ$  (range:  $0^\circ$ – $18.5^\circ$ ). None of the images in this set were used for training the CNNs or tuning the algorithm. Manual AVR (M-AVR) measurements performed on the PA radiographs were used as a baseline to compare the automatic AVR (A-AVR) ones. All M-AVR measurements were performed by a rater who had over 20 years of experience measuring scoliotic parameters manually. The rater was blinded to the A-AVR measurements.

The percentage of A-AVR measurements within clinical acceptance was calculated to determine the automatic algorithm's accuracy performance. An A-AVR measurement was deemed clinically acceptable if the measurement was within at most  $5^\circ$  of the M-AVR measurement. In addition, the circular mean absolute error (CMAE) and standard deviation of circular absolute errors (SD) were used to determine the automatic algorithm's accuracy performance. These two metrics were derived from the circular absolute error (CAE), which is defined as:

$$\text{CAE} = \arctan\left(\frac{\sin(|\theta_a - \theta_m|)}{\cos(|\theta_a - \theta_m|)}\right) \quad (3)$$

where  $\theta_a$  and  $\theta_m$  refers to the A-AVR and M-AVR measurement, respectively. Finally, our previous method [7] was applied to this 13-image EOS radiograph set to quantify the effect of the improvements from the proposed method.

**Table 1** Fivefold cross-validation results for CNN<sub>SC</sub> and CNN<sub>VB</sub>

Network	Precision	Recall	Dice coefficient
Spinal column (CNN <sub>SC</sub> )	0.959 ± 0.004	0.955 ± 0.006	0.957 ± 0.003
Vertebral body (CNN <sub>VB</sub> )	0.911 ± 0.011	0.929 ± 0.006	0.915 ± 0.006

### 3 Results and Discussion

#### 3.1 Spinal Feature Segmentation

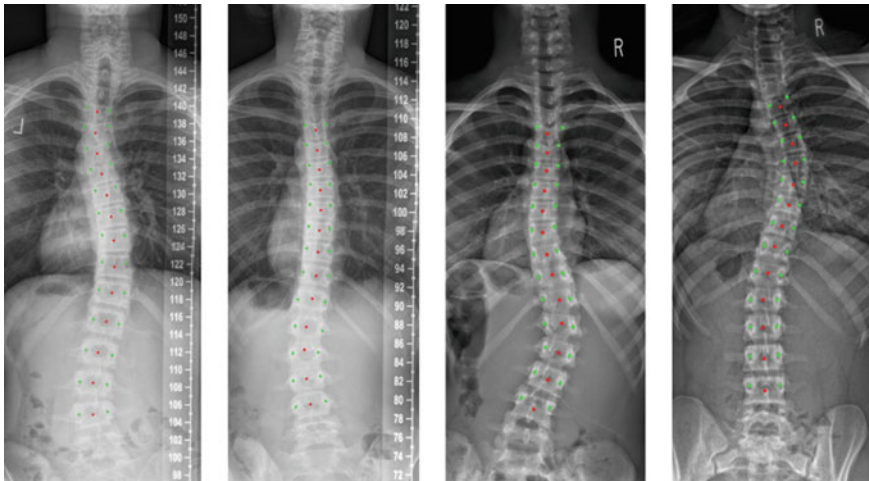
CNN<sub>SC</sub>, CNN<sub>VB</sub>, and CNN<sub>PED</sub> converged in 398, 195, and 170 epochs, respectively. Table 1 lists the fivefold cross-validation results for CNN<sub>SC</sub> and CNN<sub>VB</sub>. A Dice coefficient greater than 0.91 was achieved for all folds in both networks.

#### 3.2 AVR Measurement

In the 26-image test set, there were 338 vertebrae eligible for AVR measurement using Stokes' method (T4-L4 for each radiograph). The automatic algorithm successfully measured the AVR for all 338 vertebrae. Examples of AVR measurements on the full PA radiographs are illustrated in Fig. 4. Overall, the A-AVR measurements were 3.1° away from the M-AVR measurements on average, which is within the clinically accepted error. Also, 84% of the A-AVR measurements were within clinical acceptance of the M-AVR measurements. Table 2 lists the measurement accuracy results for different groups, according to x-ray imaging system. Looking only at EOS radiographs, our method outperformed that of Logithasan et al. [7] on the same 13-image test set by 1.4° in terms of CMAE (4.6° ± 6.2°) and 6% in terms of clinical acceptance rate (79%).

There was no notable difference between the algorithm's performance on conventional vs. EOS radiographs. There was only a 0.2° difference in terms of CMAE and a 2% difference in terms of clinical acceptance between the two groups.

Our measurement algorithm took 22.0 ± 9.9 s to measure the AVR of 13 vertebrae per radiograph. This meant that an AVR measurement for a single vertebra took 1.7 s on average. This is a significant improvement over the typical 30 s it takes to manually measure the AVR for each vertebra. The algorithm measured more quickly on radiographs taken by the EOS than the conventional system (14.3 ± 2.2 s vs. 29.7 ± 8.4 s). In contrast, the method in [7] took 109.6 ± 21.1 s to measure the 13 AVR values per radiograph in the 13-image EOS test set.



**Fig. 4** Examples of AVR measurements with pedicle centroids (green) and the center of minimum width of the vertebral body (red) labelled. The two leftmost images are from the conventional x-ray system and the two rightmost images are from the EOS x-ray system

**Table 2** Comparison results for M-AVR versus A-AVR measurements on the AVR test set

Category	% within clinical	CMAE ± SD (°)
All	84% (283/338)	3.1° ± 3.5°
Conventional	83% (140/169)	3.0° ± 3.7°
EOS	85% (143/169)	3.2° ± 3.4°

## 4 Conclusion

This manuscript reports on a fully automatic AVR measurement method on PA radiographs of children with AIS. The proposed method measured 84% of its 338 AVR measurements within the clinically accepted error and was on average only 3.1° away from the manual measurements. This accuracy, combined with the quick computation time (1.7 s per vertebra) and highly interpretable outputs, demonstrates the strong clinical feasibility of the method, giving scoliosis centers the ability of convenient AVR measurement. In addition, it could result in improved treatment outcomes and improved surgical planning, leading to a reduction in screw-related complications.

## References

1. Weinstein, S., Dolan, L., Cheng, J., Danielsson, A., Morcuende, J.: Adolescent idiopathic scoliosis. *The Lancet*. 371, 1527–1537 (2008). [https://doi.org/https://doi.org/10.1016/S0140-6736\(08\)60658-3](https://doi.org/https://doi.org/10.1016/S0140-6736(08)60658-3).

2. Lam, G., Hill, D., Le, L., Raso, J., Lou, E.: Vertebral rotation measurement: a summary and comparison of common radiographic and CT methods. *Scoliosis*. 3, 16 (2008). <https://doi.org/https://doi.org/10.1186/1748-7161-3-16>.
3. Stokes, I., Bigalow, L., Moreland, M.: Measurement of Axial Rotation of Vertebrae in Scoliosis. *Spine*. 11, 213–218 (1986). <https://doi.org/https://doi.org/10.1097/00007632-198604000-00006>.
4. Vrtovec, T., Pernus, F., Likar, B.: A review of methods for quantitative evaluation of axial vertebral rotation. *European Spine Journal*. 18, 1079–1090 (2009). <https://doi.org/https://doi.org/10.1007/s00586-009-0914-z>.
5. Zhang, J., Lou, E., Hill, D., Raso, J., Wang, Y., Le, L., Shi, X.: Computer-aided assessment of scoliosis on posteroanterior radiographs. *Med Biol Eng Comput*. 48, 185–195 (2010). <https://doi.org/https://doi.org/10.1007/s11517-009-0556-7>.
6. Ebrahimi, S., Gajny, L., Vergari, C., Angelini, E., Skalli, W.: Vertebral rotation estimation from frontal X-rays using a quasi-automated pedicle detection method. *Eur Spine J*. 28, 3026–3034 (2019). <https://doi.org/https://doi.org/10.1007/s00586-019-06158-z>.
7. Logithasan, V., Wong, J., Reformat, M., Lou, E.: Using machine learning to automatically measure axial vertebral rotation on radiographs in adolescents with idiopathic scoliosis. *Medical Engineering & Physics*. 107, 103848 (2022). <https://doi.org/https://doi.org/10.1016/j.medengphy.2022.103848>.
8. Xu, B., Wang, N., Chen, T., Li, M.: Empirical Evaluation of Rectified Activations in Convolutional Network, <http://arxiv.org/abs/1505.00853>, (2015).
9. Du, S., Zheng, N., Ying, S., You, Q., Wu, Y.: An Extension of the ICP Algorithm Considering Scale Factor. In: 2007 IEEE International Conference on Image Processing. pp. 193–196 (2007). <https://doi.org/10.1109/ICIP.2007.4379798>.

# Diagnostic Accuracy and Reliability of Deep Learning-Based Human Papillomavirus Status Prediction in Oropharyngeal Cancer



**Agustina La Greca Saint-Estevan, Chiara Marchiori, Marta Bogowicz, Javier Barranco-García, Zahra Khodabakhshi, Ender Konukoglu, Oliver Riesterer, Panagiotis Balermipas, Martin Hüßner, A. Cristiano I. Malossi, Matthias Guckenberger, Janita E. van Timmeren, and Stephanie Tanadini-Lang**

**Abstract** Oropharyngeal cancer (OPC) patients with associated human papillomavirus (HPV) infection generally present more favorable outcomes than HPV-negative patients and, consequently, their treatment with radiation therapy may be potentially de-escalated. The diagnostic accuracy of a deep learning (DL) model to predict HPV status on computed tomography (CT) images was evaluated in this study, together with its ability to perform unsupervised heatmap-based localization of relevant regions in OPC and HPV infection, i.e., the primary tumor and lymph nodes, as a measure of its reliability. The dataset consisted of 767 patients from one internal and two public collections from The Cancer Imaging Archive and was split into training, validation and test sets using the ratio 60–20–20. Images were resampled to a resolution of 2 mm<sup>3</sup> and a sub-volume of 96 pixels<sup>3</sup> was automatically cropped, which spanned from the nose until the start of the lungs. Models Genesis was fine-tuned for the classification task. Grad-CAM and Score-CAM were applied to the test subjects that belonged to the internal cohort (n = 24), and the overlap and Dice coefficients between the resulting heatmaps and the planning target volumes (PTVs) were calculated. Final train/validation/test area-under-the-curve (AUC) values of 0.9/0.87/0.87, accuracies of 0.83/0.82/0.79, and F1-scores of 0.83/0.79/0.74 were

---

A. La Greca Saint-Estevan (✉) · M. Bogowicz · J. Barranco-García · Z. Khodabakhshi · O. Riesterer · P. Balermipas · M. Guckenberger · J. E. van Timmeren · S. Tanadini-Lang  
Department of Radiation Oncology, University Hospital Zurich and the University of Zurich,  
Rämistrasse 100, 8091 Zurich, Switzerland  
e-mail: [lagreca@usz.ch](mailto:lagreca@usz.ch)

A. La Greca Saint-Estevan · E. Konukoglu  
Computer Vision Laboratory, Department of Information Technology and Electrical Engineering,  
ETH Zurich, Sternwartstrasse 7, 8092 Zurich, Switzerland

C. Marchiori · A. C. I. Malossi  
IBM Research Zurich, Säumerstrasse 4, 8803 Rüschlikon, Switzerland

M. Hüßner  
Department of Nuclear Medicine, University Hospital Zurich and the University of Zurich,  
Rämistrasse 100, 8091 Zurich, Switzerland



achieved. The reliability analysis showed an increased focus on dental artifacts in HPV-positive patients, whereas promising overlaps and moderate Dice coefficients with the PTVs were obtained for HPV-negative cases. These findings prove the necessity of performing reliability studies before a DL model is implemented in a real clinical setting, even if there is optimal diagnostic accuracy.

**Keywords** Human papillomavirus · HPV · Deep learning · Oropharyngeal cancer · OPC · Interpretability · Reliability

## 1 Introduction

The global burden of oropharyngeal cancer (OPC) has steadily increased over the last decades, reaching almost 100,000 new patients in 2020 [1]. Infection with high-risk variants of human papillomavirus (HPV) has been recognized as the leading cause of this rising incidence rate, accounting for approximately 30% of OPC cases worldwide [2]. Patients with HPV-driven tumors generally present a more favorable outcome after treatment with radiation therapy (with or without concomitant chemotherapy), with 3-year overall survival rates ranging from 82.4% for HPV-positive cases to 57.1% for HPV-negative cases [3]. The etiological and prognostic significance of HPV status has thus led to its inclusion in the American Joint Committee on Cancer (AJCC) staging guidelines for OPC and has prompted an increased interest in treatment de-escalation strategies for HPV-positive patients [4]. Current clinical standard for the determination of an HPV infection is often based on p16 staining/immunohistochemistry (IHC) due to its ease of implementation, high sensitivity, and inexpensiveness [5].

Cantrell et al. investigated the radiological differences between computed tomography (CT) images of HPV-positive and HPV-negative OPC patients [6]. Authors reported tumor exophytic characteristics, improved tumor border definition, and an increased presence of cystic nodal metastases in HPV-positive oropharyngeal carcinomas, whereas HPV-negative tumors were more likely to present muscle invasion. However, currently there is no standard, widely adopted radiological signature for the prediction of HPV status. A significant link between CT radiomic features associated to textural heterogeneity and HPV-negative tumors has been reported in several studies [7–9]. Nevertheless, radiomics involves the extraction of quantitative imaging features which are hand-engineered, which might not be relevant or well-suited for the prediction task and might be especially time-consuming.

Deep learning (DL) circumvents these disadvantages by using the entire image or region of interest as input data, and by performing automatic feature extraction and selection. Two studies were found which employed DL for the task of HPV prediction in CT of OPC [10, 11]. Both studies performed transfer learning from non-medical models and only focused on the primary tumor and its immediate regions, without taking into consideration other important regions such as the affected lymph nodes. In this study, we propose an alternative deep learning-based method for HPV status

diagnosis in CT images of OPC patients. The proposed approach performs transfer learning from Models Genesis, a publicly available 3D convolutional neural network (CNN) pre-trained on lung CT images [12]. The 3D input is selected automatically and requires little pre-processing, eliminating the need for previously delineated contours of the primary tumor, and it includes all regions from the nasal columella until the start of the lungs, capturing thus the complete spatial context of the disease. Moreover, the reliability concern regarding the black-box nature of DL methods is addressed by evaluating the ability of the classification model to perform unsupervised heatmap-based localization of the planning target volume (PTV). Our hypothesis is that a reliable model should mainly focus on the primary tumor, affected lymph nodes, and immediate surrounding regions, as these are well-known relevant areas in OPC and HPV infection.

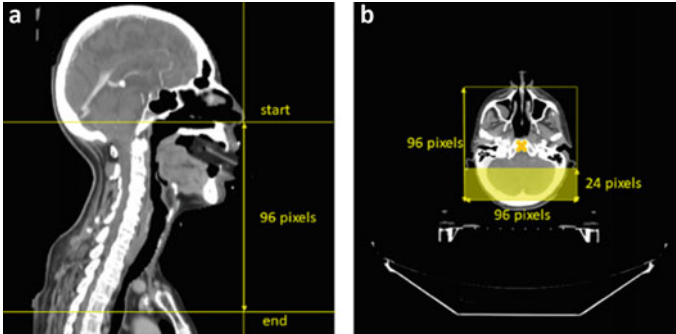
## 2 Methods

### 2.1 Study Cohort

The dataset was composed of clinical and imaging data from 767 oropharyngeal cancer patients, retrospectively collected from our institution ( $n_1 = 106$ ) following approval of the local ethical committee, and two publicly available cohorts from The Cancer Imaging Archive (TCIA) [13]: the OPC Radiomics collection ( $n_2 = 498$ ) [14] and the head and neck squamous cell carcinoma (HNSCC) collection ( $n_3 = 163$ ) [15]. Definitive chemoradiation or radiation alone were used to treat all the patients. Pre-treatment contrast-enhanced CT scans and HPV status information were available for all patients. p16 IHC was employed in most cases to determine HPV status. Contours of the PTV, defined by a radiation oncologist were available for all patients in the internal cohort. The dataset was split into training (60%), validation (20%), and test (20%) sets, so that the HPV-positive/negative ratio was kept the same in the three partitions.

### 2.2 CT Pre-processing

The pre-processing pipeline was fully automated. First, the DICOM CT scans were converted to NIfTi format and resampled to a resolution of  $2 \times 2 \times 2$  mm by interpolation. On each scan, a crop of  $96 \times 96 \times 96$  pixels was extracted, which spanned from the nasal columella to approximately the lungs' apices. The most inferior axial slice of the head which intersected with the most anterior slice of the head was selected as the starting axial slice (i.e., nose slice). The end or most caudal axial slice was selected 96 slices below the starting axial slice (Fig. 1a). The height and width dimensions of the resultant 96 axial slices were trimmed so that they would



**Fig. 1** Sub-volume selection in a head-and-neck CT scan. **a** Starting axial slice and ending slice 96 slices below. **b** Center of mass of the starting axial slice. The posterior 24 coronal slices of the sub-volume are set to background

have a square  $96 \times 96$  pixels shape (Fig. 1b). This cropping was centered around the center of mass of the starting axial slice. The posterior 24 coronal slices of the sub-volume were set to background as no relevant structures in OPC are in those regions. Additionally, CT values were clipped in the range of  $[-40, 160]$  HU and min-max normalized to have values ranging from 0 to 1. Voxels with HU values outside the selected range were set to the mean HU value to avoid sharp gradients, except for those voxels with HU values lower than  $-100$  HU, which were set to 0 after normalization.

### 2.3 Deep Neural Network: Architecture and Training

The proposed model was composed of two parts: a feature extractor and a classifier. The former consisted of the pre-trained encoder of the 3D Genesis Chest CT model [12], a 27-layer CNN of approximately 7 million parameters, whereas the latter consisted of a three-layer fully-connected network of approximately 80 thousand parameters. Adam optimizer was employed to minimize the cross-entropy loss and optimally tune model parameters. An extensive manual search was carried out to determine the following hyper-parameter combination: starting learning rate of 0.00001 with a step decay of 0.9 every two epochs; starting number of frozen layers of 15 with the unfreezing of one additional convolutional block every 100 epochs, a batch size of 4 and dropout between the fully-connected layers with a probability of 0.25. The model was trained for 500 epochs, even though early stopping was applied if the validation loss or F1-score did not improve during 50 epochs. To counteract the class imbalance, on each epoch, the model learned from a reduced, class-balanced subset of 200 subjects. Data augmentation was applied on-the-fly consisting of small random rotations and elastic deformations with a probability of 0.8, and random left-right flips with a probability of 0.5. The final model was selected based on the

best validation F1-score. The model was implemented using Keras and Tensorflow libraries.

## 2.4 Reliability Assessment

In order to assess the reliability of the proposed model, one gradient-based method, Grad-class activation map (Grad-CAM) [16], and one perturbation-based method, Score-CAM [17], were applied to the test subjects which belonged to the internal cohort ( $n = 24$ ), as the delineations of the PTVs were readily available and carried out by the same expert radiation oncologist. Grad-CAM exploits the spatial information that is maintained through the convolutional layers of the model to detect which regions in the input image are important in the classification. Typically, each feature map of the last convolutional layer is given an importance weight for a specific class  $c$ . This weight is calculated as the gradient of the output of the network for that class before softmax with respect to that feature map averaged over its width, height and depth dimensions.

Then, the output heatmap is obtained by first computing the weighted sum of the feature maps, followed by the application of ReLU and the final re-sizing of the heatmap to the input image size. Score-CAM, on the other hand, does not require the computation of gradients. The importance weights of each feature map are calculated based on the prediction score obtained after masking the input with the respective feature map and performing a forward pass. The resulting heatmap is constructed in the same manner as in Grad-CAM. After the application of the above-mentioned interpretability methods, two different threshold values were used to select the most important regions in the heatmap: the 70th and 85th percentiles. Afterwards, the overlap coefficient (OC) and the Dice similarity coefficient (DSC) between the PTVs and the thresholded heatmaps (thHMs) were calculated as follows:

$$OC = \frac{|PTV \cup thHM|}{\min\{|PTV|, |thHM|\}} \quad DSC = \frac{2*|PTV \cap thHM|}{|PTV| + |thHM|}$$

## 3 Results

### 3.1 Study Cohort

The characteristics of the patients included in each partition of the dataset are described in Table 1. There was an HPV status class imbalance of 560 positive versus 207 negative subjects (73.0% versus 27.0%).

**Table 1** Patient characteristics

	Training	Validation	Test	Total
# Patients	459	154	154	767
HPV status (pos./neg.)	335/124	113/41	112/42	560/207
Sex (M/F)	363/96	393/105	132/31	607/160
Tumor size (cm <sup>3</sup> )	27.3	26.7	30.5	27.7
Stage (7th edition))				
I–II/III–IV	27/432	12/142	16/138	55/712
Resolution (mm <sup>2</sup> )				
0.49	80	22	28	130
> 0.49, < 0.98	73	21	26	120
0.98	290	103	96	489
Other	16	8	4	28
Slice thickness (mm)				
1	80	23	29	132
2	358	126	115	599
Other	21	5	10	36

Mean values are reported for the tumor size

### 3.2 Diagnostic Accuracy and Reliability Assessment

The best model was found at epoch 220 after 6 h and 48 min of training. Final training/validation/test AUC values of 0.90/0.87/0.87, accuracies of 0.83/0.82/0.82 and F1-scores of 0.83/0.79/0.74 were obtained.

The reliability study was performed on the 24 test patients that belonged to the internal cohort. 7 out of 11 HPV-negative patients and 9 out of 13 HPV-positive patients were correctly classified by the network. The mean  $\pm$  standard deviation values for the overlap and Dice coefficients between the CAMs and the PTVs can be found in Table 2. HPV-negative samples that were correctly classified showed better overlap scores than those incorrectly classified as positives. However, no significant difference was found between true negatives and false negatives. Moreover, correctly classified HPV-positive samples showed worse mean overlap and Dice coefficients with the PTVs than those incorrectly classified. Visual inspection of the heatmaps revealed a focus on dental artifacts in those cases correctly and incorrectly classified as HPV-positive by the model (Fig. 2a). On the contrary, the heatmaps of the samples classified as HPV-negative presented a good agreement with the PTV (Fig. 2b).

**Table 2** Mean  $\pm$  standard deviation values for the overlap coefficients (OCs) and Dice similarity coefficients (DSCs) between the planning target volumes (PTVs) and the 70th-percentile- and 85th-percentile-thresholded heatmaps obtained with Grad-CAM and Score-CAM

	Thr	Grad-CAM		Score-CAM	
		OC	DSC	OC	DSC
TN (n = 7)	70th	54.3 $\pm$ 17.6	32.0 $\pm$ 11.4	<b>61.1 <math>\pm</math> 11.5</b>	30.3 $\pm$ 5.4
	85th	37.2 $\pm$ 15.5	32.0 $\pm$ 12.6	<b>38.9 <math>\pm</math> 13.2</b>	<b>30.6 <math>\pm</math> 9.0</b>
FP (n = 4)	70th	34.8 $\pm$ 6.7	21.3 $\pm$ 6.0	48.9 $\pm$ 7.6	25.8 $\pm$ 3.2
	85th	21.6 $\pm$ 4.3	18.9 $\pm$ 4.2	23.0 $\pm$ 3.8	18.8 $\pm$ 2.9
TP (n = 11)	70th	28.5 $\pm$ 10.2	17.9 $\pm$ 6.3	46.6 $\pm$ 19.9	24.6 $\pm$ 6.2
	85th	18.8 $\pm$ 8.6	15.5 $\pm$ 5.7	25.3 $\pm$ 12.6	19.7 $\pm$ 7.2
FN (n = 3)	70th	49.1 $\pm$ 10.7	<b>33.4 <math>\pm</math> 7.9</b>	56.5 $\pm$ 7.3	32.4 $\pm$ 4.9
	85th	34.2 $\pm$ 4.3	27.7 $\pm$ 3.8	31.6 $\pm$ 5.0	27.6 $\pm$ 4.6

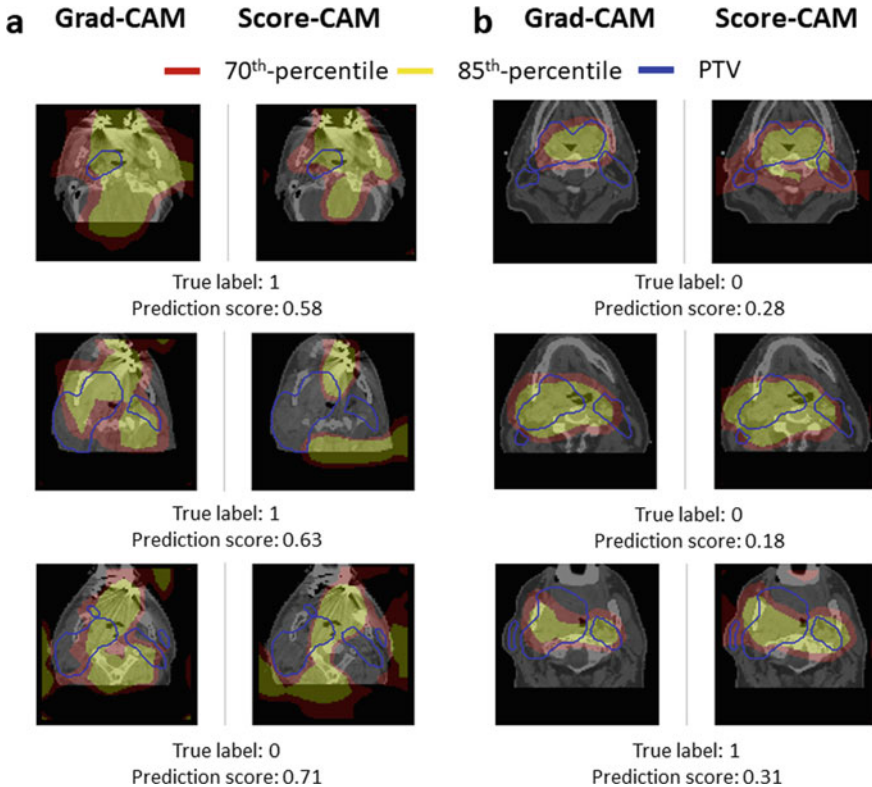
TN: true negatives; FP: false positives; TP: true positives; FN: false negatives

## 4 Discussion

Thecom goal of the study was to assess the clinical accuracy and utility of deep learning for the prediction of HPV status in CT images of OPC patients by means of deep learning, together with the assessment of the model's ability to perform unsupervised heatmap-based localization of the planning target volume as a measure of its reliability. Imaging and clinical data from 767 patients were collected from our institution and two public collections from the TCIA, and employed to fine-tune a 3D CNN for the prediction task. While the resulting model achieved excellent diagnostic performance, the reliability analysis revealed an increased focus of the model on dental artifacts in HPV-positive patients. On the contrary, promising overlap between the heatmaps and the PTVs was observed in patients classified as HPV-negative.

The diagnostic accuracy of DL for the proposed task was also explored in [10, 11]. Both studies employed transfer learning techniques from non-medical models and achieved very good performances on external patient cohorts (AUC = 0.81–0.88). Nevertheless, there were some disadvantages present in both studies: they required previously delineated tumor contours, the complete 3D spatial context of the disease was not investigated (only the primary tumor) and both lacked an evaluation of the model's reliability. In our study, we propose a fully-3D approach in which all structures located between the nose and the lungs are considered in the prediction task. Moreover, tumor or lymph node delineations are not required as the input is constructed automatically. Additionally, transfer learning is carried out from a model pre-trained on CT images, which ensures the transferability of the learned features, as opposed to transfer learning from natural images/videos datasets.

The implementation of an AI model in radiology requires a thorough assessment of its reliability and interpretability [18]. Therefore, in this study we evaluated the reliability of the model's predictions via the study of different post-hoc attribution



**Fig. 2** **a** 70th-percentile-CAM (red), 85th-percentile-CAM (yellow), and PTV (dark blue) contours for two correctly classified HPV-positive patients and one incorrectly classified HPV-negative patient. **b** 70th-percentile-CAM (red), 85th-percentile-CAM (yellow), and PTV (dark blue) contours for two HPV-negative patients correctly classified and one HPV-positive patient incorrectly classified

maps and their overlap with the PTVs. We hypothesized that a reliable prediction should focus on the primary tumor, lymph nodes and surrounding regions, as these structures are known to be affected by HPV infection. Post-hoc generated CAMs have been widely employed to interpret DL-based medical image classification tasks [19], such as detection of Parkinson’s disease [20], multiple sclerosis [21], Alzheimer’s disease [22], and COVID [23] among others, as they offer the possibility to explore the correlation between well-known disease image biomarkers and those regions relevant for the model’s prediction. Our model showed increased relevance of areas including the primary tumor and affected lymph nodes in the prediction of HPV-negative cases. However, the model failed to focus on the PTV in HPV-positive cases and focused instead on dental artifacts. This learnt correlation between the presence of HPV infection and dental implants could be a result of different OPC patient profiles: HPV-negative patients usually have a history of tobacco, alcohol and other substances consumption, which is frequently linked to poor oral hygiene

and care [24]. On the contrary, HPV-positive patients are associated to an increased number of sexual partners, with no association to tobacco or other substances abuse. It is likely that this group is more careful about dental care and has more artificial implants. This example shows the importance of performing a reliability analysis to find model limitations before they are introduced in the clinic.

Several limitations were found in the proposed study. Firstly, the scarcity of labeled data potentially hampered the generalization capability of the model. Furthermore, head-and-neck CT scans are very frequently subject to artifacts from dental implants, which cause streak patterns that may affect the study of the primary tumor and other regions of interest and as observed in this study, may confound any automatic image analysis tool. The inclusion of magnetic resonance and positron emission tomography images, which have been successfully employed to predict HPV status [25], could lead to a more accurate and explanatory diagnostic tool. Another potential drawback encountered was the use of p16 IHC for ground truth data labeling, as this technique might lead to moderate false-positive rates [27]. mRNA-in situ hybridization-labelled data should hence be considered to rigorously assess the clinical validity of the proposed method. Finally, further studies on external validation cohorts should be carried out to validate the reported findings.

## 5 Conclusion

DL achieved excellent performance in HPV status diagnosis in CT of OPC. However, our reliability analysis showed an increased relevance of regions with dental artifacts for the prediction of HPV-positive cases, whereas a good agreement with the PTV was observed in HPV-negative cases. These findings prove the necessity of performing reliability studies before a deep learning model is implemented in a real clinical setting, even if there is an optimal diagnostic accuracy.

**Funding** This work was supported by the Swiss National Science Foundation (310030 173303), the EMDO foundation and the SPHN project IMAGINE.

## References

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*. 71, 209–249 (2021).
2. de Martel, C., Plummer, M., Vignat, J., Franceschi, S.: Worldwide burden of cancer attributable to HPV by site, country and HPV type. *Int J Cancer*. 141, 664–670 (2017).
3. Ang, K.K., Harris, J., Wheeler, R., Weber, R., Rosenthal, D.I., Nguyen-Tân, P.F., Westra, W.H., Chung, C.H., Jordan, R.C., Lu, C., Kim, H., Axelrod, R., Silverman, C.C., Redmond, K.P., Gillison, M.L.: Human papillomavirus and survival of patients with oropharyngeal cancer. *N Engl J Med*. 363, 24–35 (2010).



4. Masterson, L., Moualed, D., Liu, Z.W., Howard, J.E.F., Dwivedi, R.C., Tysome, J.R., Benson, R., Sterling, J.C., Sudhoff, H., Jani, P., Goon, P.K.C.: De-escalation treatment protocols for human papillomavirus-associated oropharyngeal squamous cell carcinoma: A systematic review and meta-analysis of current clinical trials. *European Journal of Cancer*. 50, 2636–2648 (2014).
5. Fauzi, F.H., Hamzan, N.I., Rahman, N.A., Suraiya, S., Mohamad, S.: Detection of human papillomavirus in oropharyngeal squamous cell carcinoma. *J Zhejiang Univ Sci B*. 21, 961–976 (2020).
6. Cantrell, S.C., Peck, B.W., Li, G., Wei, Q., Sturgis, E.M., Ginsberg, L.E.: Differences in imaging characteristics of HPV-positive and HPV-Negative oropharyngeal cancers: a blinded matched-pair analysis. *AJNR Am J Neuroradiol*. 34, 2005–2009 (2013).
7. Bogowicz, M., Riesterer, O., Ikenberg, K., Stieb, S., Moch, H., Studer, G., Guckenberger, M., Tanadini-Lang, S.: Computed Tomography Radiomics Predicts HPV Status and Local Tumor Control After Definitive Radiochemotherapy in Head and Neck Squamous Cell Carcinoma. *Int J Radiat Oncol Biol Phys*. 99, 921–928 (2017).
8. Lee, J.Y., Han, M., Kim, K.S., Shin, S.-J., Choi, J.W., Ha, E.J.: Discrimination of HPV status using CT texture analysis: tumour heterogeneity in oropharyngeal squamous cell carcinomas. *Neuroradiology*. 61, 1415–1424 (2019).
9. Mungai, F., Verrone, G.B., Pietragalla, M., Berti, V., Addeo, G., Desideri, I., Bonasera, L., Miele, V.: CT assessment of tumor heterogeneity and the potential for the prediction of human papillomavirus status in oropharyngeal squamous cell carcinoma. *Radiol Med*. 124, 804–811 (2019).
10. Lang, D.M., Peeken, J.C., Combs, S.E., Wilkens, J.J., Bartzsch, S.: Deep Learning Based HPV Status Prediction for Oropharyngeal Cancer Patients. *Cancers (Basel)*. 13, 786 (2021).
11. La Greca Saint-Estevan, A., Bogowicz, M., Konukoglu, E., Riesterer, O., Balermipas, P., Guckenberger, M., Tanadini-Lang, S., van Timmeren, J.E.: A 2.5D convolutional neural network for HPV prediction in advanced oropharyngeal cancer. *Comput Biol Med*. 142, 105215 (2022).
12. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models Genesis. *Medical Image Analysis*. 67, 101840 (2021).
13. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository | SpringerLink, <https://link.springer.com/article/https://doi.org/10.1007/s10278-013-9622-7>, last accessed 2021/02/10.
14. Kwan, J.Y.Y., Su, J., Huang, S.H., Ghorraie, L.S., Xu, W., Chan, B., Yip, K.W., Giuliani, M., Bayley, A., Kim, J., Hope, A.J., Ringash, J., Cho, J., McNiven, A., Hansen, A., Goldstein, D., De Almeida, J.R., Aerts, H.J., Waldron, J.N., Haibe-Kains, B., O’Sullivan, B., Bratman, S.V., Liu, F.-F.: Data from Radiomic Biomarkers to Refine Risk Models for Distant Metastasis in Oropharyngeal Carcinoma, <https://wiki.cancerimagingarchive.net/x/XAQQAg>, (2019).
15. Grossberg, A., Mohamed, A., El Halawani, H., Bennett, W., Smith, K., Nolan, T., Chamchod, S., Kantor, M., Browne, T., Hutcheson, K., Gunn, G., Garden, A., Frank, S., Rosenthal, D., Freymann, J., Fuller, C.: Data from Head and Neck Cancer CT Atlas, <https://wiki.cancerimagingarchive.net/x/CoFyAQ>, (2017).
16. 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: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017).
17. Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X.: Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 111–119. IEEE, Seattle, WA, USA (2020).
18. Salahuddin, Z., Woodruff, H.C., Chatterjee, A., Lambin, P.: Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine*. 140, 105111 (2022).
19. Magesh, P., Myloth, R., Tom, R.: An Explainable Machine Learning Model for Early Detection of Parkinson’s Disease using LIME on DaTSCAN Imagery. *Computers in Biology and Medicine*. 126, 104041 (2020).

20. Eitel, F., Soehler, E., Bellmann-Strobl, J., Brandt, A.U., Ruprecht, K., Giess, R.M., Kuchling, J., Asseyer, S., Weygandt, M., Haynes, J.-D., Scheel, M., Paul, F., Ritter, K.: Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *Neuroimage Clin.* 24, 102003 (2019).
21. Böhle, M., Eitel, F., Weygandt, M., Ritter, K.: Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification. *Frontiers in Aging Neuroscience.* 11, (2019).
22. Panwar, H., Gupta, P.K., Siddiqui, M.K., Morales-Menendez, R., Bhardwaj, P., Singh, V.: A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. *Chaos Solitons Fractals.* 140, 110190 (2020).
23. Gillison, M.L., D'Souza, G., Westra, W., Sugar, E., Xiao, W., Begum, S., Viscidi, R.: Distinct Risk Factor Profiles for Human Papillomavirus Type 16–Positive and Human Papillomavirus Type 16–Negative Head and Neck Cancers. *JNCI: Journal of the National Cancer Institute.* 100, 407–420 (2008).
24. Fujima, N., Andreu-Arasa, V.C., Meibom, S.K., Mercier, G.A., Truong, M.T., Sakai, O.: Prediction of the human papillomavirus status in patients with oropharyngeal squamous cell carcinoma by FDG-PET imaging dataset using deep learning analysis: A hypothesis-generating study. *European Journal of Radiology.* 126, 108936 (2020).
25. Wang, H., Zhang, Y., Bai, W., Wang, B., Wei, J., Ji, R., Xin, Y., Dong, L., Jiang, X.: Feasibility of Immunohistochemical p16 Staining in the Diagnosis of Human Papillomavirus Infection in Patients With Squamous Cell Carcinoma of the Head and Neck: A Systematic Review and Meta-Analysis. *Front. Oncol.* 0, (2020).
26. Reyes, M., Meier, R., Pereira, S., Silva, C.A., Dahlweid, F.-M., Tengg-Kobligk, H. von, Summers, R.M., Wiest, R.: On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiology: Artificial Intelligence.* 2, e190043 (2020).

# Optimizing the Illumination of a Surgical Site in New Autonomous Module-based Surgical Lighting Systems



Andre Mühlenbrock, René Weller, and Gabriel Zachmann

**Abstract** Good illumination of the surgical site is crucial for the success of a surgery—yet current, typical surgical lighting systems have significant shortcomings, e.g. with regard to shadowing and ease of handling. To address these shortcomings, new lighting systems for operating rooms have recently been developed, consisting of a variety of swiveling light modules that are mounted on the ceiling and controlled automatically. For such a new type of lighting system, we present a new optimization pipeline that maintains the brightness at the surgical site as constant as possible over time and minimizes shadows by using depth sensors. Furthermore, by performing simulations on point cloud recordings of nine real abdominal surgeries, we demonstrate that our optimization pipeline is capable of effectively preventing shadows cast by bodies and heads of the OR personnel.

**Keywords** Surgical lighting · Optimal illumination · Depth sensors

## 1 Introduction

Although good illumination of the surgical site—i.e. the wound—is so important for the success of an operation, existing solutions, e.g. conventional surgical lighting system (SLS) or head lamps, have major disadvantages. In the case of SLS, the main drawback is the shadowing by OR personnel around the table that makes illumination of an OR wound difficult and requires frequent manual readjustments of the SLS. Head lamps, on the other hand, are strenuous to wear for long periods of time and require the wearer to assume a certain head posture.

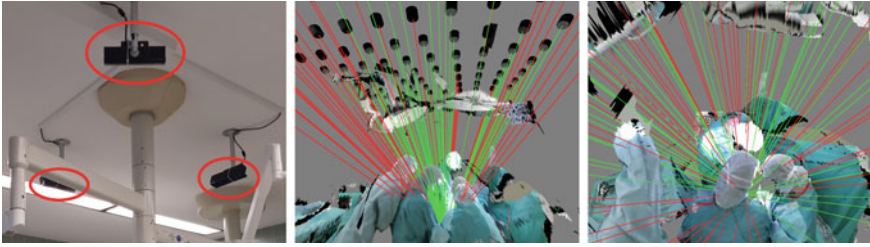
---

A. Mühlenbrock (✉) · R. Weller · G. Zachmann  
Faculty of Computer Science, University of Bremen, Bremen, Germany  
e-mail: [muehlenb@cs.uni-bremen.de](mailto:muehlenb@cs.uni-bremen.de)

R. Weller  
e-mail: [weller@cs.uni-bremen.de](mailto:weller@cs.uni-bremen.de)

G. Zachmann  
e-mail: [zach@cs.uni-bremen.de](mailto:zach@cs.uni-bremen.de)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
R. Su et al. (eds.), *Medical Imaging and Computer-Aided Diagnosis*, Lecture Notes  
in Electrical Engineering 810, [https://doi.org/10.1007/978-981-16-6775-6\\_24](https://doi.org/10.1007/978-981-16-6775-6_24)



**Fig. 1** Left: The Kinect cameras we used to record nine real open abdominal surgeries in an operating room. Center and right: The resulting point cloud recordings against which we can perform ray tests

To solve the problems of these SLS, new surgical lighting systems have been developed to automatically prevent shadows and to keep the brightness in the surgical site at the desired level constantly over time and as evenly distributed as possible over the area. These new surgical lighting systems do not consist of two or three conventional large lighting systems, but of a large number of small swiveling light modules placed at the ceiling that are automatically rotated and intensity-controlled with the aid of a central control computer. Recent examples are the Optimus ISE Celestial™ Surgical Lighting System,<sup>1</sup> and the lighting system developed in the SmartOT research project.<sup>2</sup>

In this paper, we present a novel optimization pipeline for such lighting systems based on multiple depth sensors such as Microsoft's Kinect. By simulating the optimization with point cloud recordings of nine real open abdominal surgeries (see Fig. 1), we compare different parameters and fitness scores with respect to the illumination they create at a virtual surgical site.

## 2 Related Work

In today's operating rooms, SLS are commonly used for traditional open surgery. However, they still come with some disadvantages: According to Knulst et al.[1], conventional surgical lights are readjusted every 7.5 min on average to provide appropriate illumination for the surgical site. In addition, surgeons and other OR personnel often saw a need for improvement in lighting intensity, shadowing, illumination of deep wounds, and the handling of such lights. Curlin et al. [2] also elaborates on the advantages and disadvantages of surgical lights and other common lighting systems, including head lights, lighted retractors, and operating microscopes, none of which meet all lighting needs.

<sup>1</sup> See <https://www.optimus-ise.com/>.

<sup>2</sup> See <https://www.smart-ot.de/>.

As an approach to improve the handling of conventional SLS, Dietz et al. [3] suggest to use a gesture control for brightness and color temperature instead of using a control panel, which is usually located high up on the SLS. An attempt to also address the problem of manual repositioning and alignment of conventional surgical lights was provided by Teuber et al. [4], in which three motor-driven surgical lights automatically position themselves so that shadows are avoided. This optimization was further optimized in [5]. We have discussed developing these ideas further and implementing a similar motor-driven approach, but have rejected it due to several drawbacks, including the expected noise and the danger in terms of collisions with OR personnel.

In novel lighting concepts for operating rooms, as in the SmartOT project, a variety of small lighting modules are proposed that are placed on the ceiling and control themselves to automatically generate optimal illumination at the site and avoid shadowing. Recently, an optimization procedure was presented in [6] to position the light modules of such lighting systems on the ceiling with the help of point cloud recordings in such a way that the most satisfactory illumination is theoretically reachable during the entire surgery.

Nevertheless, to the best of our knowledge, methods to optimize the intensity of light modules at the runtime of the surgery for this new type of surgical lighting system have not been presented or evaluated in the literature up to this point.

### 3 Implementation

In this section, we present our optimization pipeline (Sect. 3.3) as well as the specific optimization of the intensities of individual light modules (Sect. 3.4). For understanding, we briefly discuss the different types of shadows beforehand in Sect. 3.1 and describe our surgical site model for enabling the illumination of deep wounds in Sect. 3.2.

#### 3.1 Occluder Types

The shadows in surgeries can be divided into two categories: On the one hand, there are shadows caused by hands and OR instruments, where occluders—the hands and instruments—are very close to the site. These shadows are difficult to compensate for by an autonomous shadow management because the lights in question change very quickly due to the fast movements and the short distance of the occluder to the surgical site. The most time, even all the lights cast a shadow for these type of occluders. In new module-based lighting concepts, these shadows can be compensated by distributing as many light modules as possible over a large area which are used simultaneously.

On the other hand, there are shadows caused by the heads and bodies of OR personnel: for these type of occluders, only some lamps cast shadows at the surgical

site simultaneously, since the head and body of individual persons are usually located to the side of the surgical site and their distance to the site is greater. In this section, we mainly focus on preventing this type of shadowing.

### 3.2 Representation of the Surgical Site

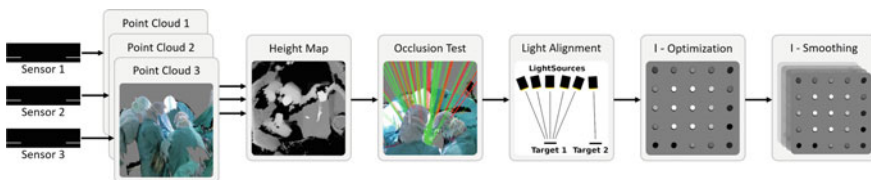
In order to illuminate narrow, deep surgical sites, the site is modeled using a virtual cylinder-shaped tube, which can be placed, rotated and scaled in diameter and depth. By not just testing for occlusions by the point cloud geometry but also against this tube, we can ensure that only light modules are used which are able to illuminate the site in depth when this is required. This virtual surgical site model is visualized in Fig. 3.

### 3.3 Optimization Pipeline

Our pipeline (see Fig. 2) starts with the depth images from multiple depth cameras as input. The cameras are placed on the ceiling between the lamps. By using the camera intrinsics and extrinsics parameters, a point cloud is generated in camera space, registered to each other and transformed into world space—in our case the OR room. In case of the point cloud recordings of the nine surgery used for evaluation, we used a lattice registration procedure [7] to extrinsically calibrate the depth cameras.

In order to efficiently test for occlusions, we first transform the separate point clouds of multiple depth sensors into a common geometric datastructure, which is a height map that stores the height from the ground in the 2 m x 2 m area around the operating table. In a first step, we remove occlusions of the first type according to Sect. 3.1, i.e. hands and instruments close to the site. This can be easily achieved by simply removing all points of the point cloud within a fixed radius  $r$  around the site (we used  $r = 0.3$  m).

By projecting a ray into this height map and iterate over the resulting line, we can efficiently calculate for each light module whether there is an occluding object



**Fig. 2** The pipeline starts with the input of the depth sensors and ends with the output of the parameters to control the light modules, described in detail in Sect. 3.3

in the path that would lead to shadows at the site. To test whether a ray is blocked by the surgical site model (see Sect. 3.2), a simple ray-plane intersection test can be performed where the distance between the intersection point and plane midpoint is tested against the tube radius. By testing multiple rays per light module which are starting at different positions on the luminous surface and run to different positions in the site, we moreover calculate a floating point value  $v_i$  that indicates how much a light module  $i$  is blocked by the geometry. To reduce sensor noise, we filter this value using a 1D Kalman filter.

Next, the light modules in the pipeline are assigned to a light target and rotated accordingly so that they are aligned with it. This is done according to the desired setting defined by the OR staff. Finally the optimization and smoothing of the intensities of the light modules takes place which is described in Sect. 3.4.

### 3.4 Light Intensity Optimization

**Requirements** Since we want to optimize the intensity  $I^i$  of every light module  $i$  in such a way that the illuminance  $E_v$  at the surgical site is as constant as possible and close to the desired value  $E_{v\text{pref}}$ , we need be able to calculate what illumination  $E_v^i$  is produced by a single light module  $i$  at the site.

In order to be able to calculate which intensity values  $I^i$  produce which illumination  $E_v^i$  at the center of the site for any one of the modules of the light module array, we assume that for an arbitrary chosen, but fixed distance  $d_{\text{Norm}}$  and a perpendicular incidence of light, a mapping function  $f$  is known that maps the intensity  $I^i$  the light module  $i$  is driven with to illumination  $E_{v\text{Norm}}^i$ :

$$E_{v\text{Norm}}^i = f(I^i) \tag{1}$$

Given the distance  $d^i$  from a light module  $i$  to the center of the surgical site, the virtual site surface normal  $\mathbf{n}$  and the light vector  $\mathbf{l}^i$  of light module  $i$ , we approximate the luminance  $E_v^i$  as follows:

$$E_v^i = E_{v\text{Norm}}^i \cdot \left(\frac{d_{\text{Norm}}}{d^i}\right)^2 \cdot (-\mathbf{l}^i \cdot \mathbf{n}) \tag{2}$$

Note that we assume that the illuminance of the used light modules decreases approximately quadratically over distance. The term  $(-\mathbf{l}^i \cdot \mathbf{n})$ , on the other hand, describes the decrease in luminance when the surface on which the same amount of light is incident increases due to a tilt—similar to Lambert’s cosine law.

#### I-Optimization

The optimization approach can be summarized by the following two steps: In the first step, the light modules are sorted according to their suitability to illuminate the site well. In the second step, each non-occluded light module is assigned a certain

amount of light until the target brightness is reached, starting with the most suitable light module.

To sort the light modules, we implemented two different scores. The first score is the perpendicularity of the inverse light vector  $-\mathbf{l}$  to the site surface with the surface normal  $\mathbf{n}$ :

$$s_{Perpendicular}^i = (-\mathbf{l}_i \cdot \mathbf{n}) \quad (3)$$

The second score we have implemented counts the number of last consecutive frames in which the floating-point number value  $v$  filtered with the 1D Kalman filter (see Sect. 3.3) was greater than 0.9:

$$s_{SuccessiveUnblocked}^i = \#\text{Consecutive frames with } v^i \geq 0.9 \quad (4)$$

After sorting the light modules by their presumed ability to illuminate the site well, we calculate the maximum illuminance  $E_{v_{\text{Max}}}$  the system is able to provide at the center of the surgical site. To do this, we use a simple heuristic: The maximum illuminance  $E_{v_{\text{Max}}}^i$  a light module  $i$  can produce at the center of the site is multiplied by the relative number of unblocked light rays from that light module  $i$  to the site and then summed up over all light module:

$$E_{v_{\text{Max}}} = \sum^i E_{v_{\text{Max}}}^i \cdot \text{unblockedRays}(i) \quad (5)$$

Before we iterate over the list of sorted light modules, we define the ratio which describes how much illuminance is preferred ( $E_{v_{\text{Pref}}}$ ) compared to the illuminance  $E_{v_{\text{Max}}}$  the system is actually able to provide using all unblocked light modules:

$$\omega = \min\left(1, \frac{E_{v_{\text{Pref}}}}{E_{v_{\text{Max}}}}\right) \quad (6)$$

In addition, we define a variable  $E_{v_{\text{Rem}}}$  that is decreased over time and describes which illuminance is still needed to reach the preferred illuminance  $E_{v_{\text{Pref}}}$ . Accordingly, it is initialized with the preferred illuminance:

$$E_{v_{\text{Rem}}} \leftarrow E_{v_{\text{Pref}}} \quad (7)$$

Finally, we iterate over the list of sorted light modules and calculate the intensity  $I^i$  with which each light module  $i$  should be driven. In order to investigate how the number of simultaneous lights used affects the characteristics of the illumination, we have adapted our optimization method to be configurable by a floating-point light spread parameter  $\alpha$ , which specifies whether as few optimal and unblocked light modules as possible should be used ( $\alpha = 0.0$ ), or whether the desired brightness should be achieved by using all the unblocked light modules ( $\alpha = 1.0$ ):

$$I^i = I_{\text{Max}}^i (1 - \alpha) + I_{\text{Max}}^i \cdot \omega \cdot \alpha \quad (8)$$



Moreover, we calculate the illuminance  $E_v^i$  expected to be achieved at the site for the light module  $i$  by using Eqs. (1) and (2):

$$E_v^i = f(I^i) \cdot \left(\frac{d_{\text{Norm}}}{d_i}\right)^2 \cdot (-\mathbf{l}_i \cdot \mathbf{n}) \quad (9)$$

In the case that this value  $E_v^i$  is lower than the remaining required illumination, i.e.  $E_v^i \leq E_{v_{\text{Rem}}}$ , the illumination of this light module  $i$  is subtracted from the remaining needed illumination:

$$E_{v_{\text{Rem}}} \leftarrow E_{v_{\text{Rem}}} - E_v^i \quad (10)$$

In the case that the illumination of light module  $i$  would exceed the remaining needed illumination, i.e.  $E_v^i > E_{v_{\text{Rem}}}$ , we recalculate the intensity with which the light module should be driven:

$$I^i \leftarrow I^i \cdot \frac{E_{v_{\text{Rem}}}}{E_v^i} \quad (11)$$

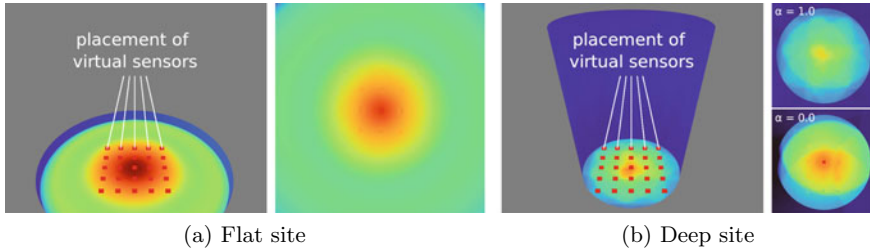
as well as we set the remaining needed illumination to zero and stop the iteration at that light. The intensity of all other light modules is left at zero.

**I-Smoothing** Since we do not want the light to react immediately to every movement around the site, as this might distract surgeons, we perform a temporal smoothing of the intensity values  $I^i$  for every light  $i$ . We do this by simply blending the already smoothed intensities  $I_{\text{Smoothed}}^{i,t-1}$  of the previous frame  $t - 1$  with the optimal luminous power  $I^{i,t}$  of the current frame  $t$  by using a blending value  $\gamma \in (0, 1]$ :

$$I_{\text{Smoothed}}^{i,t} = I_{\text{Smoothed}}^{i,t-1} \cdot (1 - \gamma) + I^{i,t} \cdot \gamma \quad (12)$$

**Remarks** Currently, we shoot multiple rays from a single light module to different points at the surgical site to calculate the unblocked amount of light  $v_i$  for each light module  $i$  (see description of  $v_i$  in Sect. 3.3), but for intensity optimization, we only consider the illumination at a single point, i.e. the center of the surgical site. However, it would also be possible to optimize the illumination not only for a single point but for the whole surface area: This might be particularly useful if the emission characteristics of the light module are distributed unevenly over the surface (e.g. for cost reasons of the installed LED as seen in Fig. 3).

Nevertheless, such an optimization causes some problems: On the one hand, currently, our site model is only very coarse, mainly, because of the limited resolution and the fixed viewing angle of the depth sensors that are not able to capture the complex geometry and details of real world sites. Moreover, the emission characteristics of the *actual physical light modules*—e.g. beam angle—cannot be changed. Consequently, the only way to compensate for uneven emission patterns remains the continuous rotation of the light modules.



**Fig. 3** Visualization of the sensors placed onto (a) an almost flat surgical site and (b) a deep surgical site with a depth of 10 cm and a diameter of 7.5 cm. Here, the illumination is color coded using the Google turbo color map

But even if optimization over the entire area currently seems to make little sense due to these problems, our pipeline as a whole is prepared to handle this as we are able to estimate the illumination in multiple points at a site.

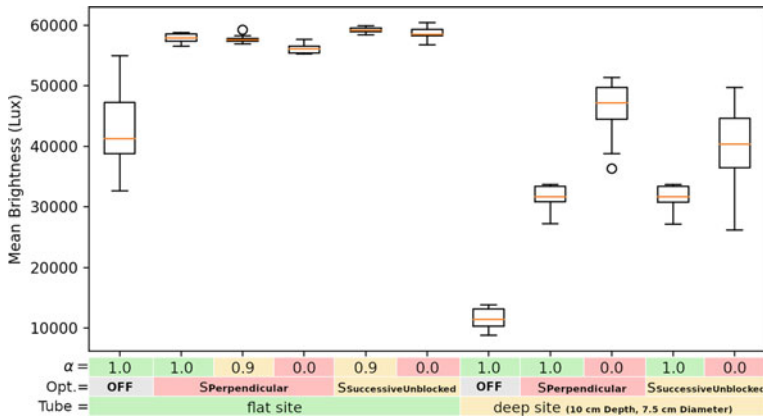
## 4 Results

In this section, we evaluate our optimization pipeline in a simulation on point cloud recordings of nine real abdominal surgeries. The methodology of the evaluation is presented in Sect. 4.1. Since the best possible light in the surgical situs is generally assumed to be (a) as free of shadows as possible and (b) should not change visibly as much as possible to avoid interference, we examine the quality of illumination with respect to these aspects in Sects. 4.2 and 4.3.

### 4.1 Methods

For the evaluation, we used point cloud recordings of nine real abdominal surgeries taken at Pius-Hospital Oldenburg, Germany (see Fig. 1). The setup of the evaluated virtual lighting system was as follows: We used  $7 \times 8$  light modules at a height of 2.5 m arranged in a grid with a spacing of 36 cm  $\times$  35 cm. The preferred illumination  $E_{v_{\text{pref}}}$  was set to 80 klx. Single light modules were able to generate almost 50 klx at the site center at a distance of 1.9 m when driven at maximum intensity. We placed  $5 \times 5$  sensors on an area of 5 cm  $\times$  5 cm in the virtual site and simulated the brightness at 60 Hz using an illumination profile of the actual planned light modules in the SmartOT prototype, generated and provided by Qioptiq Photonics GmbH & Co. KG (see Fig. 3).

For our measurements, we chose a representative scene of 1:30 min from each OR recording and placed the virtual site to the position where the real surgical site was in that specific recording. In order not to be biased by the choice of scene, we decided



**Fig. 4** Mean brightness: Brightness averaged over sensors and over time (n = 9 surgery sections)

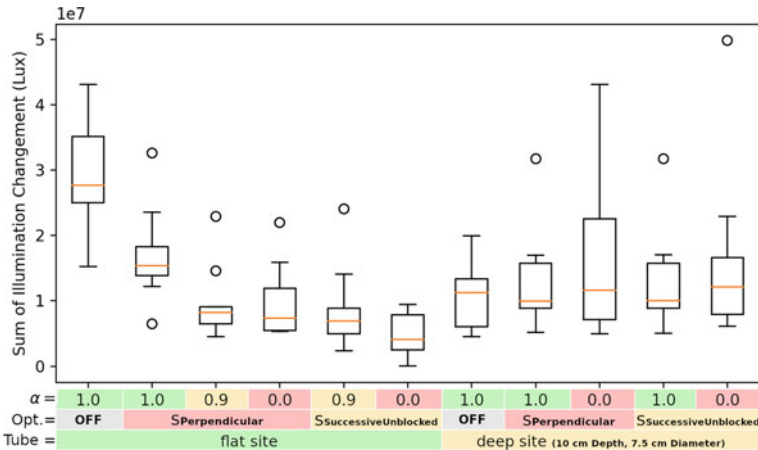
to use a scene in the middle of each recording, i.e. at exactly 2h after the start of the recording. Moreover, we discarded the first 10s for warm-up of the Kalman and smoothing filter.

Finally, in the evaluation we examined how different parameters affect the lighting properties, which are (a) the parameter  $\alpha$  presented in Sect. 3.4, which specifies the amount of simultaneous used light modules, (b) the score functions presented in Sect. 3.4, where ‘OFF’ represents no optimization and no response of the light modules to occluding geometry and (c) the usage on a flat wound (without a shadow casting tube) or a deep surgical site, see Fig. 3.

## 4.2 Shadow Reduction

First, we examined the average brightness and plotted it in Fig. 4 as this indicates the amount of shadowing. The average brightness in the flat site with optimization is 56.2klx–59.2klx depending on the setting (compared to 43.0klx without optimization), which is very close to the expected optimum without shadows with about 57–62klx depending on the position of the site (keep in mind that the preferred illumination of 80klx refers just to the maximum value in the center of the site). Moreover, with the flat site, the settings regarding brightness have practically no impact.

However, in case of a deep site, our results show that they will be illuminated very low without optimization with an average of 11.4klx—after all, the light from most lamps does not penetrate at all. While the brightness of both optimization scores is identical with  $\alpha = 1.0$  and is 31.4klx, since simply all available lamps are used, one sees that  $S_{\text{Perpendicular}}$  with an average of 45.6klx performs slightly better than  $S_{\text{SuccessiveUnblocked}}$  with 39.4klx, which might be explained by the fact that the light



**Fig. 5** Changes over time: Sum of illumination changes between adjacent frames, summed over all sensors (n = 9 surgery sections)

modules which can shine most vertically into the site and cause less shadows at the edge of the tube always tend to be selected.

### 4.3 Temporal Brightness Distribution

Comparing the brightness changes over time, it is noticeable that there are significantly more changes at the site without optimization than with activated optimization (see Fig. 5). However, except for a large  $\alpha = 1.0$ , where more changes occur over time than with  $\alpha \leq 0.9$ , the optimization settings have little effect on the overall rate of change when optimization is activated.

## 5 Conclusions and Future Works

We have presented a simple optimization algorithm for optimizing the illumination of a surgical site for new module-based lighting systems with a large number of swiveling automatically controlled light modules which are using depth sensors. We have investigated the influence of individual optimization parameters, namely, the number of simultaneously used lights and the depth of a virtual surgical wound with respect to the average brightness and changes in brightness. Finally, with our simulation, we were able to show that automatic optimization of intensity is a very effective means of preventing shadows and providing uniform illumination of the site over time.

In future work, we will evaluate the new lighting concept with the presented optimization within a real prototype and conduct a user study with active surgeons performing a task similar to actual operation. In this user study, we will also compare the performance to conventional SLS. Finally, we would like to consider the whole site *area* instead of only the site *center* for the optimization (see Sect. 3.4).

**Acknowledgements** This work was partially funded by BMBF grant 13GW0264D.

## References

1. Arjan J Knulst et al. “Indicating shortcomings in surgical lighting systems”. en. In: *Minim Invasive Ther Allied Technol* 20.5 (Nov. 2010), pp. 267–275.
2. Jahnvi Curlin and Charles K Herman. “Current State of Surgical Lighting”. en. In: *Surg J (N Y)* 6.2 (June 2020), e87–e97.
3. Armin Dietz et al. “Contactless Surgery Light Control based on 3D Gesture Recognition”. In: *GCAI 2016. 2nd Global Conference on Artificial Intelligence*. Ed. by Christoph Benzmlüller, Geoff Sutcliffe, and Raul Rojas. Vol. 41. EPiC Series in Computing. 2016, pp. 138–146.
4. Jörn Teuber et al. “Autonomous Surgical Lamps”. In: *Jahrestagung der Deutschen Gesellschaft für Computer- und Roboterassistierte Chirurgie (CU-RAC)*. Bremen, Germany, Sept. 2015.
5. Jörn Teuber et al. “Optimized Positioning of Autonomous Surgical Lamps”. In: *Proceedings of the SPIE Medical Imaging Conference*. Orlando, FL, United States of America: SPIE, Feb. 2017.
6. Andre Mühlenbrock et al. “Optimizing the arrangement of fixed light modules in new autonomous surgical lighting systems”. In: *Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling*. Ed. by Cristian A. Linte and Jeffrey H. Siewerdsen. Vol. 12034. International Society for Optics and Photonics. SPIE, 2022, pp. 491–499.
7. Andre Mühlenbrock et al. “Fast, accurate and robust registration of multiple depth sensors without need for RGB and IR images”. In: *The Visual Computer* (May 2022). ISSN: 1432-2315.

# An Eye-Tracking Based Machine Learning Model Towards the Prediction of Visual Expertise for Electrocardiogram Interpretation



Mohammed Tahri Sqalli , Dena Al-Thani , Mohamed B. Elshazly ,  
Mohammed Al-Hijji , Alaa Alahmadi , and Yahya Sqalli Houssaini 

**Abstract** The electrocardiogram, known as the ECG or EKG, is considered among the mostly used medical diagnostic tests worldwide. Despite the test's prevalence in the healthcare sector, there still exist gaps into training medical practitioners become skilled and efficient ECG interpreters. Moreover, this also brings the challenge of assessing the expertise of those practitioners. This is primarily due to the difficulty of assessing visual expertise. Visual expertise is the skill of interpreting images relating to a certain technical field. Due to the limited quantitative research methodologies that could not capture this subtle skill during the previous two decades, a limited number of models are being conceptualized and assessed. In addition, automated ECG interpretation models based on artificial intelligence are still not accurate enough to be fully deployed in the medical field. This therefore leaves only one choice, which is to focus on improving on methodologies to train and assess medical practitioners' visual expertise. This approach will contribute towards increasing the accuracy

---

M. T. Sqalli (✉)

Division of Engineering, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

Department of Economics, School of Foreign Services, Georgetown University in Qatar, Doha, Qatar

e-mail: [mts517@nyu.edu](mailto:mts517@nyu.edu)

D. Al-Thani

Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

M. B. Elshazly

Department of Cardiology and Electrophysiology, Medical University of South Carolina, North Charleston, SC, USA

M. Al-Hijji

Interventional & Structural Cardiology Division, Heart Hospital at Hamad Medical Corporation, Doha, Qatar

A. Alahmadi

College of Computer Science and Engineering, Taibah University, Yanbu, Saudi Arabia

Y. S. Houssaini

Department of Medicine, Mohammed V University Faculty of Medicine and Pharmacy, Rabat, Morocco

of ECG interpretations within medical institutions by forming competent medical staff. In this paper, we present a road map for the development of an eye-tracking based machine learning model that leads towards the prediction of visual expertise within medical practitioners. To develop the model, we built on top of our previously conducted research that aimed at understanding the differences in visual patterns within medical practitioners with different expertise levels. The developed model could predict the expertise level of the ECG interpreter with an accuracy of 94.08%. This is thanks to the eye movement patterns of the participant.

**Keywords** Electrocardiogram · ECG · ECG interpretation · Eye-tracking · Human-computer interaction

## 1 Introduction

### 1.1 Background

The electrocardiogram, abbreviated as the ECG or EKG, “is a graph that represents the electrical activity of the human heart” [1]. The ECG is referenced as the main-stream initial medical diagnostic test to support the clinical decision making process regarding the patient’s heart. Thanks to its unified structure, it englobes the diagnostic of both mild and urgent heart conditions [2]. The ECG is the most-used medical test globally with 300 million ECGs done yearly in the United States alone [2]. This raise in demand over the past decade for ECG interpretation has pushed the need for more expertise capable of an accurate and immediate interpretation. This is with the aim to decrease the number of avoidable cardiac deaths. Moreover, as the price and effort for taking one has decreased over the years, patients are more referred to getting an ECG for an effective diagnosis [3]. The spotlight is therefore directed towards medical practitioners to handle this imminent and continuous flow of uninterpreted cases. During these circumstances, clinical staff may make conscious or non-conscious compromises during the interpretation process. These compromises range from clinical staff not being fully trained to carry out the task of interpretation, to interpreters not having the full access to the patient’s clinical information and history [4]. The shortcoming for this strategy may result in dangerous consequences [3].

The repercussions of an inaccurate ECG interpretation are detrimental. They range from avoidable cardiac deaths, to unnecessary expensive costs that burden the patient. In an investigative report conducted by Mele et al. [5] tens of millions of dollars are poured annually into malpractice payouts, in addition to the 100,000 patients who die annually due to an avoidable cardiac death. Moreover, another study that assessed ECG interpretation competency in cardiology residents describes that 58% of imminent potential life-threatening ECGs are missed by residents [6]. 60% of those residents claim to be convinced that they perform a correct interpretation [6].

These findings urge for the improvement of ECG interpretation. This is by developing more effective training programs, as well as having the necessary tools that enable the identification of where do medical practitioners lack crucial ECG interpretation expertise.

## ***1.2 Related Works***

The literature gap that this paper addresses is the lack of use of machine learning to understand and predict the visual expertise of medical practitioners interpreting ECGs. In addition to the scarcity of literature that used eye tracking for ECG interpretation, each study of the three available studies has looked at eye tracking data to solve one small portion of a bigger problem which is the ECG interpretation, but not the problem as a whole. Examples of that would be Bond et al. [7] that adopted eye tracking in order to only understand how the collected eye-gaze data may be used in order to unveil insights about how expert annotators proceed with an ECG interpretation. This study was restricted to only the quantification of experts' interpretation of an ECG, while not considering the potential of applying machine learning on the eye tracking data. The following studies by Davies et al. [8, 9] proposed a qualitative study that followed a similar eye tracking study design. Under the aim of better understanding the applied cognitive processes that expert interpreters refer to when interpreting ECGs, Davies et al. [8] referred to interviewing and surveying rather than quantitatively analyzing the eye tracking data. Finally, the most recent study is the one conducted by Wu et al. [10] where the aim was to understand visual expertise within medical practitioners. The study is designed to be a qualitative one using eye tracking as a supporting analytical element. The aim was to highlight the differences in the cognitive approaches to ECG interpretation between medical students, emergency medicine (EM) residents, and EM attending physicians using both interviews and eye tracking experiments. This work is the first one that applies machine learning on eye tracking data to predict the visual expertise levels and patterns for medical practitioners.

The objective of this paper is to design a machine learning model able to predict the visual behavior, represented by the fixation's duration over the ECG lead, for medical practitioners with different expertise levels. The aim is to reach an accuracy of over 90%. In the upcoming sections, we describe the dataset for eye-tracking data along with its pre-processing methodologies (Sect. 2). We then describe the data processing methodologies through the feature engineering process (Sect. 3). This enables us to design and craft a model able to predict the fixation duration across all of the electrocardiogram leads for the medical practitioners. Subsequently, we present this model in the results section (Sect. 4). Finally, we conclude the paper (Sect. 5).



## 2 The Dataset

### 2.1 Overview

The dataset [11], open-sourced through PhysioNet.org [12], was collected across two phases for an exploratory eye tracking study. The first phase had as a goal unveiling the ECG interpretation dynamics for medical students [13]. The second phase extended the first one by including, in addition to medical students, medical practitioners [14]. These practitioners had varying expertise levels in ECG interpretation as well as different roles in the medical sector. They deal with ECGs mostly daily while practicing in the hospital. The primary aim behind the collection of this dataset was to use it for uncovering insights to chart key best practices as well as common mistakes in the ECG interpretation process [15]. The following aim behind the dataset collection, which relates to this paper, was to find innovative solutions to improve the ECG interpretation training programs for both medical students and practitioners [16]. This may be through the inclusion of eye-tracking as well as machine learning as supporting technologies. The collected eye tracking data was used as the primary quantitative data for the below studies that performed a quantitative analysis on the medical students' data [13], as well as the remaining medical practitioners' data [14].

### 2.2 Data Collection Method

The dataset was generated through the collection of eye-tracking data using a Tobii Pro X2-60 eye tracker device and iMotions version 8.1 software [17]. Medical practitioners' eye movements were sampled with a frequency 60Hz ( $\pm 1$  Hz). Gaze and fixations data related to the practitioner's eye movements were recorded following their eyes' micro-saccades in terms of milliseconds. According to iMotions, "micro-saccades represent movements that are shorter in the distance that is covered compared to normal saccades, at around 15 arcminutes" [18]. The fixations algorithm used by iMotions is the I-VT (Velocity-Threshold Identification) Fixation Filter algorithm [19] where the gap fill-in option/interpolation is disabled. The noise reduction is also disabled. The following are the fixations filter parameters used, with the window length being 20 ms, and the velocity threshold being equal to 30°/s. The merger of adjacent fixations is disabled and the discard of short fixations option is also disabled. Regarding the data collection setting, the eye movement data was collected in real-time while the participants were performing the ECG interpretation in the experiment setting. The dataset collected includes a total of 630 different ECG interpretations from 63 unique medical practitioners. These practitioners belong to five expertise categories. Each practitioner or student interpreted ten different ECGs each. Moreover, Each participant in the experiment was given a limited time of 30 s to interpret the ECG. More details about the data collection methodology, the ECGs, and processes used in the experiment can be found through the dataset's repository [11].

### 2.3 Ethics

Prior to the commencement of the experiment and data collection, the study was granted institutional review board approval from the ethical board of the Qatar Biomedical Research Institute at Hamad bin Khalifa University under the research protocol number QBRI-IRB-2020-01-009. All the necessary procedures and approvals were granted before the start of the experiment. The Institutional review board approval guarantees that all study methods were conducted following the guidelines and recommendations of international regulatory agencies. All the participants in the study gave written informed consent.

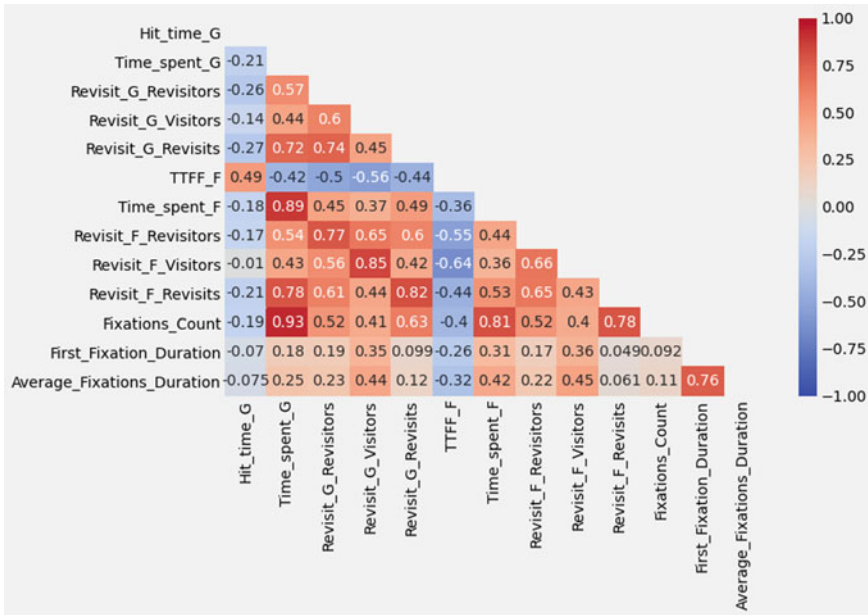
## 3 Data Processing

Prior to the initiation of the process of features exploration, data cleaning and pre-processing was done. During the pre-processing phase, explained in details in the dataset's repository description [11], eye tracking data was calculated according to each lead among the 12 leads that constitute the electrocardiogram. In addition, lines where the eye tracking data was missing were omitted. Finally all the participants were standardised by being assigned to the compelling category according to a survey they filled prior to the eye tracking experiment.

The goal in this section is to identify the relationship among all the collected eye tracking features within the set of medical students and practitioners. These features are described in the open-sourced dataset published in the PhysioNet repository [11]. For this purpose, we use the Pearson's correlation coefficient [20].

We use the eye tracking features, defined according to the dataset collection methodologies section in the dataset repository[11], to generate Pearson's correlation matrix. The generated matrix sheds light on important insights that contribute towards completing the understanding of how medical practitioners interpret an ECG. The matrix is calculated using the grid-based Areas of Interest (AOI) distribution. An area of interest represent an important area in the electrocardiogram to be interpreted. Within this AOI, all the eye tracking metrics within this area are calculated with respect to the other areas highlighted on the ECG. Our publications [13, 14] demonstrate in details the process of choosing AOIs and calculating the eye tracking metrics within these AOIs. Generally, an AOI represents an ECG lead. The generated correlation matrix highlights the intensity of the correlation among all eye tracking features respective of all AOIs. It also serves to identify the top correlated features that guide the analysis of the data.

The correlation matrix, depicted in Fig. 1, follows a colour scheme commonly known as the hot/cold colour scheme. The "hot" squares in the matrix denote a strong positive correlation among two eye tracking features, while the "cold" squares denote the opposite. In Fig. 1, most of the eye tracking features have positive correlations among each others. Some features show a strong correlation among each others



**Fig. 1** Correlation matrix showcasing the relationship among the eye tracking features for the medical practitioners

whether the nature of this correlation is positive or negative. An important observation in the correlation matrix is the existence of two vertical and one horizontal “cold” lines in the features matrix. These lines showcase the eye tracking features that are negatively correlated with the time to first fixation and with the hit time. These two features (TTFF and Hit Time) as defined in the dataset repository [11] hold small values (in milliseconds) whenever the eye tracking behavior of the participant is significant. However, it is the opposite for all the other features. Hence, the positive correlation among each other, and negative correlation among the rest of the features. This indicates that the ECG lead corresponding to a specific AOI that is fixated the earliest is potentially to be fixated more.

From the correlation matrix, we deduce the top three correlated features of the collected eye tracking data among medical students and practitioners. These are the fixations count, the time spent fixating and the fixations re-visitations with Pearson correlation coefficients equal to  $r = 0.97$ ,  $r = 0.87$ , and  $r = 0.82$  respectively. These features’ definitions can be found in the dataset’s repository [11]. These features will be the foundation of the features’ engineering process for designing the best predictive model.

## 4 Results

We explore to what extent are medical students’ interpretation patterns predictable. We attempt the comparison of different machine learning models in order to find the optimal model for the type of data presented above. Table 1 summarizes the classification reports of four models in addition to the baseline model not being trained. The train-test split is 75–25%, and the target is similar across all models for the time spent fixating across different ECG leads. The following describes the parameters of the best-performing machine learning model. It is noteworthy that the reason why advanced models like support vector machine and convolutional neural network did not perform well in term of accuracy, is mainly due to the shortage of the available eye tracking data in the dataset. We therefore refer to the best performing model among the ones tested in Table 1 to predict the visual ECG interpretation behavior of medical practitioners. We use a linear regression machine learning model to attempt to predict to what extent are visual patterns of medical practitioners predictable and thus their visual expertise level. The collected dataset contains around 298,589 raw observations. These observations contain the gaze and fixation data, along with some noisy data due to poor lighting conditions and poor eye-calibrations. Filtering the raw data results in 45,917 fixations. These fixations are then clustered based on their *X,Y* coordinates on the ten ECG images, and based on their time sequence. This final step then results in obtaining 15,373 clusters to be analyzed. Although the sampling rate of the used eye-tracker collects a considerable amount of information, it is noteworthy that the results presented are more of a proof of concept rather than a robust predictive algorithm. This is mainly because the selected population sample is small. Moreover, the sample of ECGs explored is small as well compared to the complexity and the varying nuances of ECGs found in the clinical practice. It is also worth reminding that the aim of the predictive algorithm is to predict the medical students’ interpretation patterns and not the ECG interpretation itself. The filtered dataset, the raw dataset, and the python notebook are available upon request.

**Table 1** Average results for the classification report of different machine learning models applied to the ECG interpretation dataset

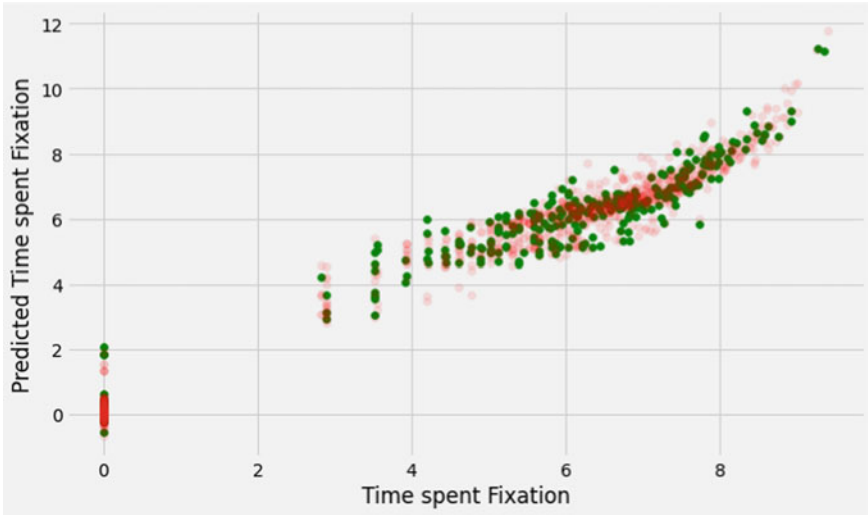
Machine learning model	Precision	Recall	F1 score	Average accuracy
Baseline	0.21	1.00	0.35	0.21
Support vector machine	0.67	0.70	0.68	0.65
Convolutional neural network	0.61	0.40	0.48	0.43
Logistic regression	0.88	0.90	0.83	0.86
Linear regression	0.94	0.88	0.95	0.92

We use the ridge regression algorithm to power our predictive machine learning model. Linear regression refers to a linear approach to model the relationship between a dependent variable and one or more independent variables [11]. The other variables explored through the correlation matrix in Fig. 1 are fed to the algorithm as independent variables. Moreover, we use one-hot encoding to encode the non-numeric variables like the ECG diagnosis and the AOI/lead name. This would enable answering the following question. Given the ECG diagnosis, and given the lead area in that ECG, is there a possibility of predicting the fixation duration of participants in that specific area of the ECG. Using an 75% train, 25% test split, the predictive algorithm reaches a predictive accuracy of 94.08%. This was done under a 5-fold cross validation by randomly shuffling the eye tracking ECG interpretation results for the best performing model that predicts the interpreter's visual expertise represented by the fixations data. we have thus obtained five AUCs. The average AUC for this was derived to be around 0.94. Figure 2 summarizes the model predictive results.

Figure 2a compares the actual interpreter's time spent fixation data points (in seconds) with the model predicted time. Since the training data is significantly larger than the testing data according to the train test split, the transparency of the training data in red is lowered by 90%. This is to allow the testing data in green to show better. Figure 2 shows that there is a linearity and a correlation between the training data and test data results, which explains the obtained accuracy score. It is also worth mentioning that the data clusters around two areas, the area which the actual time spent fixating is bigger than 2 s, and the area in which the time spent fixating is around 0 s. This is due to the scarcity or the non-availability of fixations in some AOIs. This result confirms the bi-modality in some ECGs in the fixation duration distribution found in our previous research [13].

Figure 2b displays the top 20 learned variables used by the machine learning model to make a prediction. Along with each variable, there is a coefficient assigned by the algorithm. The coefficients play a vital role in making the predictions. They indicate the direction of the relationship between a predictor variables and the predicted variable. A positive variable coefficient sign, labeled in Fig. 2b with the green colour, indicates that as the predictor variable increases, the predicted variable also increases. However, a negative coefficient sign, labeled in Fig. 2b in red indicates that as the predictor variable increases, the predicted variable decreases. We notice that the first four features in the list of top 20 used variables, are all linked to the gaze and fixations participants visitors and re-visitors. This is a logical behavior, since as the number of participant visitors to the AOI increases, there is a strong probability that this AOI contains a clue that catches the attention of the observer. Hence, the individual time spent fixating at that AOI for a sample participant will increase. This might also indicate that participants will re-visit this area more as shown in the top 11th feature. Other than the quantitative variables, the qualitative variables also have a direct influence on the time spent fixating at a specific AOI. All the E All these features and others along with their coefficients weave the complex structure of the interpretation behavior among medical students.

(a) Actual interpreters' time spent fixating at AOI vs. predicted time spent fixating



(b) predictive model learned top 20 learned parameters for prediction

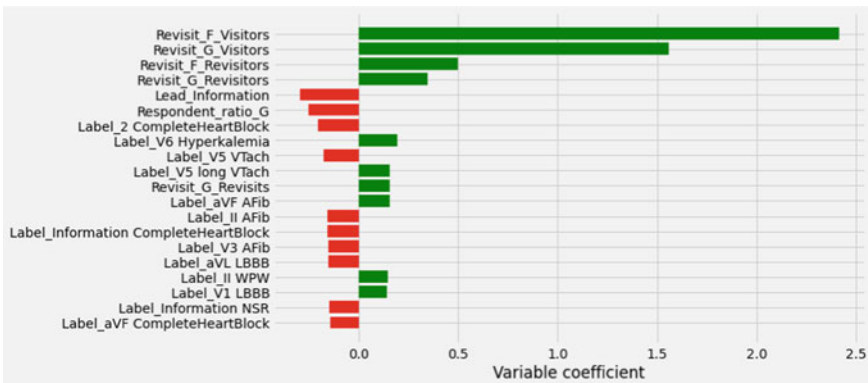


Fig. 2 Predictive model results

## 5 Conclusion

We presented a roadmap for an eye-tracking based machine learning model aiming towards the prediction of ECG interpretation behavior of medical professionals. This is with the end-goal to assess and better train medical practitioners reach accurate and efficient ECG interpretations. Throughout this paper, we described the dataset for eye-tracking data along with its pre-processing methodologies. We then described our feature engineering process. This enabled us to design and craft a model able to predict the fixation duration across all of the electrocardiogram leads for the medical practitioners. The best-performing model used linear regression to reach an

accuracy percentage of 92%. It also referred to a number of different eye-tracking features, mainly areas of interest, fixations and gaze counts and duration. This will enable us in the future works to explore the remaining strongly correlated features analyzed to predict not only the accuracy of interpretation as one static output, but also the dynamism of the scan-paths followed by these practitioners before giving their diagnosis answer. It is important to note that the model presented represents only a proof of concept demonstrating that the ECG interpretation within the medical practitioners population is predictable using machine learning techniques. This leads us to the limitations of this work. Our work is limited by the scarcity of the eye tracking data collected, which may have led to over-fitting. Despite the high sampling rate of the eye-tracker used in the experiment, machine learning models remain data-hungry. Thus, the insufficiency of the collected eye tracking data. Moreover, the sample of ECGs explored is small as well compared to the complexity and the varying nuances of ECGs found in the clinical practice.

**Acknowledgements** The authors would like to thank all the volunteering participants who contributed with their electrocardiogram interpretations.

## References

1. A. Davies and A. Scott, *Starting to Read ECGs: A comprehensive Guide to Theory and Practice*. Springer London, 2015. [Online]. Available: <https://doi.org/10.1007/978-1-4471-4965-1>
2. J. Eldridge, D. Richley, S. Eggett, C. and Baxter, S. Blackman, C. Breen, C. Brown, B. Campbell, C. Cox, J. Hutchinson, and C. Rees, E. and Ross. (2017) Clinical guidelines by consensus recording a standard 12-lead electrocardiogram an approved methodology by the society for cardiological science and technology.
3. M. P. Turakhia, M. Desai, H. Hedlin, A. Rajmane, N. Talati, T. Ferris, S. Desai, D. Nag, M. Patel, P. Kowey, J. S. Rumsfeld, A. M. Russo, M. T. Hills, C. B. Granger, K. W. Mahaffey, and M. V. Perez, "Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The apple heart study," *American Heart Journal*, vol. 207, pp. 66–75, jan 2019. [Online]. Available: <https://doi.org/10.1016%2Fj.ahj.2018.09.002>
4. P. O'Meara, G. Munro, B. Williams, S. Cooper, F. Bogossian, L. Ross, L. Sparkes, M. Browning, and M. McClounan, "Developing situation awareness amongst nursing and paramedicine students utilizing eye tracking technology and video debriefing techniques: A proof of concept paper," *International Emergency Nursing*, vol. 23, no. 2, pp. 94–99, Apr. 2015. [Online]. Available: <https://doi.org/10.1016/j.ienj.2014.11.001>
5. P. Mele, "Improving electrocardiogram interpretation in the clinical setting," *Journal of Electrocardiology*, vol. 41, no. 5, pp. 438–439, Sep. 2008. [Online]. Available: <https://doi.org/10.1016/j.jelectrocard.2008.04.003>
6. K. E. O'Brien, M. L. Cannarozzi, D. M. Torre, A. J. Mechaber, and S. J. Durning, "Training and assessment of ECG interpretation skills: Results from the 2005 CDIM survey," *Teaching and Learning in Medicine*, vol. 21, no. 2, pp. 111–115, Apr. 2009. [Online]. Available: <https://doi.org/10.1080/10401330902791255>
7. R. Bond, T. Zhu, D. Finlay, B. Drew, P. Kligfield, D. Guldenring, C. Breen, A. Gallagher, M. Daly, and G. Clifford, "Assessing computerized eye tracking technology for gaining insight into expert interpretation of the 12-lead electrocardiogram: an objective quantitative approach," *Journal of Electrocardiology*, vol. 47, no. 6, pp. 895–906, Nov. 2014. [Online]. Available: <https://doi.org/10.1016/j.jelectrocard.2014.07.011>

8. A. Davies, "Examining expertise through eye movements: A study of clinicians interpreting electrocardiograms," Ph.D. dissertation, The University of Manchester, 2018.
9. A. Davies, G. Brown, M. Vigo, S. Harper, L. Horseman, B. Splendiani, E. Hill, and C. Jay, "Exploring the relationship between eye movements and electrocardiogram interpretation accuracy," *Scientific Reports*, vol. 6, no. 1, Dec. 2016. [Online]. Available: <https://doi.org/10.1038/srep38227>
10. W. Wu, A. K. Hall, H. Braund, C. R. Bell, and A. Szulewski, "The development of visual expertise in ECG interpretation: An eye-tracking augmented re situ interview approach," *Teaching and Learning in Medicine*, pp. 1–20, Dec. 2020. [Online]. Available: <https://doi.org/10.1080/10401334.2020.1844009>
11. M. Tahri Sqalli, D. Al-Thani, M. Elshazly, and M. Al-Hijji, "Eye tracking dataset for the 12-lead electrocardiogram interpretation of medical practitioners and students," 2022. [Online]. Available: <https://physionet.org/content/eye-tracking-ecg/1.0.0/>
12. A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13), circulation Electronic Pages: <http://circ.ahajournals.org/content/101/23/e215.full> PMID:1085218; <https://doi.org/10.1161/01.CIR.101.23.e215>
13. M. T. Sqalli, D. Al-Thani, M. B. Elshazly, and M. Al-Hijji, "Interpretation of a 12-lead electrocardiogram by medical students: Quantitative eye-tracking approach," *JMIR Medical Education*, vol. 7, no. 4, p. e26675, Oct. 2021. [Online]. Available: <https://doi.org/10.2196/26675>
14. M. T. Sqalli, D. Al-Thani, M. B. Elshazly, M. Al-Hijji, A. Alahmadi, and Y. S. Hous-saini, "Understanding cardiology practitioners' interpretations of electrocardiograms: An eye-tracking study," *JMIR Human Factors*, vol. 9, no. 1, p. e34058, Feb. 2022. [Online]. Available: <https://doi.org/10.2196/34058>
15. M. T. Sqalli, D. Al-Thani, M. B. Elshazly, M. Al-Hijji, and Y. S. Houssaini, "The journey towards an accurate electrocardiogram interpretation: An eye-tracking study overview," in *2021 8th International Conference on Behavioral and Social Computing (BESC)*. IEEE, Oct. 2021. [Online]. Available: <https://doi.org/10.1109/besc53957.2021.9635168>
16. M. T. Sqalli, D. Al-Thani, M. B. Elshazly, and M. Al-Hijji, "A blueprint for an AI & AR-based eye tracking system to train cardiology professionals better interpret electrocardiograms," in *Persuasive Technology*, ser. Lecture notes in computer science. Cham: Springer International Publishing, 2022, pp. 221–229.
17. "Tobii pro x2-60," Feb 2021. [Online]. Available: <https://imotions.com/hardware/tobii-x2-60/>
18. B. Farnsworth. (2020, may) 10 most used eye tracking metrics and terms. [Online]. Available: <https://imotions.com/blog/10-terms-metrics-eye-tracking/>
19. iMotions. (2020, may) Visualizing eye tracking data. [Online]. Available: <https://help.imotions.com/hc/en-us/articles/360010834259-Visualizing-Eye-Tracking-Data>
20. K. Pearson. (2020, may) Spss tutorials: Pearson correlation. [Online]. Available: <https://libguides.library.kent.edu/SPSS/PearsonCorr>



# Synthetic Data as a Tool to Combat Racial Bias in Medical AI: Utilizing Generative Models for Optimizing Early Detection of Melanoma in Fitzpatrick Skin Types IV–VI



Daniel Kvak, Eva Březinová, Marek Biroš, and Robert Hrubý

**Abstract** Assistive tools to aid in skin cancer detection are experiencing an unprecedented rise with the accessibility of robust and accurate deep learning models. However, in the present applications, only a negligible number of dermatology images come from patients with Fitzpatrick skin types IV–VI, representing brown, dark brown or black skin, respectively. In this study, we demonstrate the utilization of Zero-Shot Text-to-Image autoregressive models to generate synthetic medical data for improved balance in training CAD classification models with minimized racial bias. Synthetically generated images of skin lesions were assessed by an experienced dermatologist using the ABCD rule and differential diagnostics, and subsequently validated using a pre-trained ResNet50V2 multi-class classification model.

**Keywords** Autoregressive models · Computer-aided diagnosis · Deep learning · Generative adversarial networks · Melanoma · Synthetic data · Zero-shot learning

## Abbreviations

AI	Artificial Intelligence
CAD	Computer-aided diagnosis system
CNN	Convolutional neural network

---

D. Kvak (✉) · M. Biroš · R. Hrubý  
Carebot s.r.o., Prague, Czech Republic  
e-mail: [daniel.kvak@carebot.com](mailto:daniel.kvak@carebot.com)

E. Březinová  
Faculty of Medicine, Masaryk University, Brno, Czech Republic

First Department of Dermatovenerology, St. Anne's University Hospital, Brno, Czech Republic

M. Biroš  
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

R. Hrubý  
Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University, Prague, Czech Republic

GAN	Generative adversarial networks
AKIEC	Actinic keratosis and intraepithelial carcinoma
BCC	Basal cell carcinoma
BKL	Benign keratosis
DF	Dermatofibroma
MEL	Melanoma
NV	Melanocytic nevus
SCC	Squamous cell carcinoma
VASC	Vascular lesion

## 1 Introduction

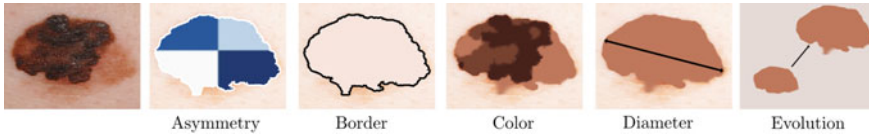
Melanoma is considered the most aggressive form of skin cancer [1]. Given the similar shape of benign and malignant findings, doctors spend considerably more time diagnosing these lesions [2]. Currently, the malignancy assessment is mainly performed by invasive histological examination of the suspicious lesion [3]. The development of an accurate classifier can reduce and monitor the negative effects of skin cancer and improve patient survival rates [4]. Currently, however, Fitzpatrick skin types IV–VI, which correspond to brown and dark brown or black skin, respectively, represent a small percentage of dermatology images available [5].

## 2 Background

Despite the fact that people with Fitzpatrick skin types IV–VI are less likely to develop melanoma, they face a higher risk of mortality because of delayed detection and treatment [6, 7]. Very often, skin cancer in these patients is diagnosed at a more advanced stage, making treatment difficult [8]. Although the incidence of melanoma in the Fitzpatrick skin type 0–III population has increased by almost 20% in the last 20 years alone [9], an epidemiological review published by the American Academy of Dermatology showed that the 5-year survival rate in the Fitzpatrick skin type IV–VI population is 70%, which is significantly lower than the Fitzpatrick skin type 0–III population (92%) [6, 7].

### 2.1 Medical Examination

Melanoma represents the most dangerous form of skin cancer. Although it is less common than other types of cancer, it often metastasizes and spreads to other parts of the body faster [10]. Melanoma develops from neoplastic proliferation of melanocytes, however, the pathophysiology is not yet clearly understood [11]; several pathogenetic



**Fig. 1** The ABCDE rule for skin cancer detection

**Table 1** Dataset characteristics indicating data imbalance across race and ethnicity taken from Jain et al. [15]

Race and ethnicity	Development set	Validation set A	Validation set B
American Indian or Alaska Native	142 (0.1%)	42 (0.1%)	9 (0.9%)
Asian	1775 (11.0%)	473 (12.6%)	97 (10.1%)
Black or African American	1087 (6.8%)	229 (6.1%)	61 (6.3%)
Hispanic or Latino	7044 (43.7%)	1631 (43.4%)	409 (42.5%)
Native Hawaiian or Pacific Islander	224 (1.4%)	61 (1.6%)	19 (2.0%)
White	5475 (34.0%)	1175 (31.3%)	329 (34.2%)
Not specified	367 (2.2%)	145 (3.9%)	39 (4.0%)

mechanisms of the development are hypothesized. Malignant melanoma is predominantly observed on the skin, but it can also develop in ears, eyes leptomeninges, and oral or genital mucosa [12]. Melanoma originates not only on sun-exposed skin, where the main pathogenetic factor is considered to be UV radiation, but also in body parts that are otherwise relatively protected from radiation [10, 13]. If melanoma is suspected, the lesion with the surrounding skin or mucosa is biopsied, followed by histological examination [12] (Fig. 1).

## 2.2 CAD Examination

Machine learning-driven systems used in healthcare are statistical models created mostly from retrospectively collected and demographic-specific data, with the aim of providing outputs that serve as a basis for decision-making [14]. Their use in allocating resources and determining access to health services has quickly become commonplace [14].

Among the most discussed commercial applications for more effective skin lesion detection is Google Health’s research [15, 16], which recognizes 26 common skin diseases, representing 80% of cases seen in primary care, while also providing secondary prediction for 419 skin diseases (Table 1).

Detection and monitoring of malignant skin tumors and benign moles is a particularly challenging problem due to the general uniformity of large skin lesions [17], the fact that the skin lesions do not differ much in appearance [18], and the relatively small amount of existing datasets [19–21]. While there has been a significant amount of attention paid to the design and validation of commercial [22–24] and academic [25, 26] CAD prediction systems, the problem of insufficient availability of suitable, high-quality and, above all, comprehensive datasets is often overlooked [27].

Furthermore, there are still no globally standardized methods for verifying the reliability of AI models: there is no single answer to the regulation or accountability of artificial intelligence in healthcare [28]. Each depends on the clinical application, intended use, use instructions, product claims, etc. [29]. Currently, there are complex software regulations that can be adapted to AI, for example in the FDA medical device assessment methodology [30]. An audit conducted before deploying a system into clinical practice could significantly improve outcomes for patients [30]. Although pre-deployment audits do not directly address underlying structural issues, they can identify and reduce the harmful effects of bias and add evidence that relying on data and models to make important decisions automatically makes outcomes more fair and honest [31].

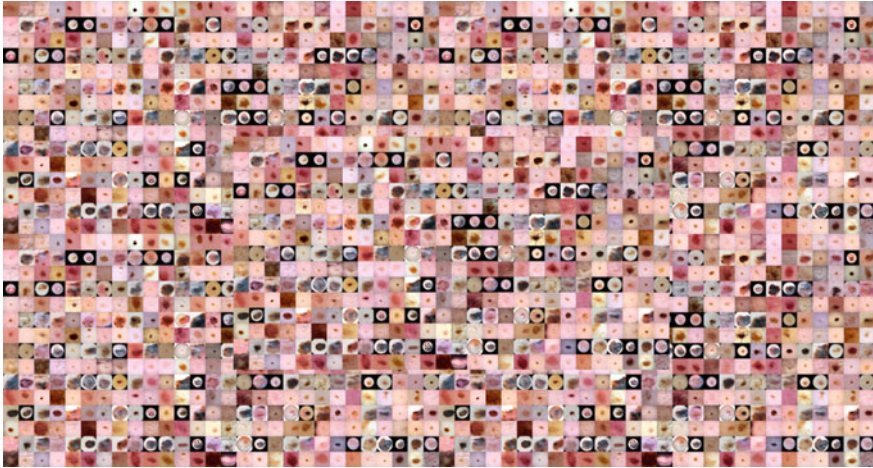
### 3 Related Works

While in less regulated fields, we can encounter the use of synthetic data at all levels of R&D [32, 33], in medical applications, they can find their use mainly in the context of internal validation of data-driven software applications [34, 35]. Data augmentation techniques (inverting, cropping, enlarging, or distorting images) can be applied to medical images, but are insufficient to produce new and original data [36].

#### 3.1 *Generative Adversarial Networks*

Among the widespread modalities are magnetic resonance imaging (MRI) [37] or computed tomography (CT) [38]. However, the path from synthetic data to clinical practice is still unclear. The use of these data as one of the sources for training machine learning models raises many questions. Nevertheless, in recent years, we can observe the initial deployment of similar generative models in drug design or protein engineering [39, 40].

A study by Li et al. [20], which addresses the possibility of using synthetic data in clinical R&D, dates back to 2016, two years after original study by Goodfellow et al. [41] which introduced generative adversarial networks (GAN). In their study, Li et al. focused on the problem of limited data availability of clinical data in biomedical research, validating synthetically generated data using a pre-trained CNN-based multi-class classifier.



**Fig. 2** Example of generated samples from Limeros et al. [42]

The lack of large open medical databases and the subsequent possibility of using GANs for generating dermatoscopic images of various skin lesions is addressed in Limeros et al. [42]. As input dataset for training GAN, they used the widely-adopted ISIC collection.<sup>1</sup> The synthetic images, shown on Fig. 2, are then evaluated through latent space exploration and embeddings projection, demonstrating the authenticity and generalization of the trained GANs [42]. From a medical perspective, the artificially generated data were also assessed by two dermatologists who failed to identify features in the images that indicated that they were real or synthetic images.

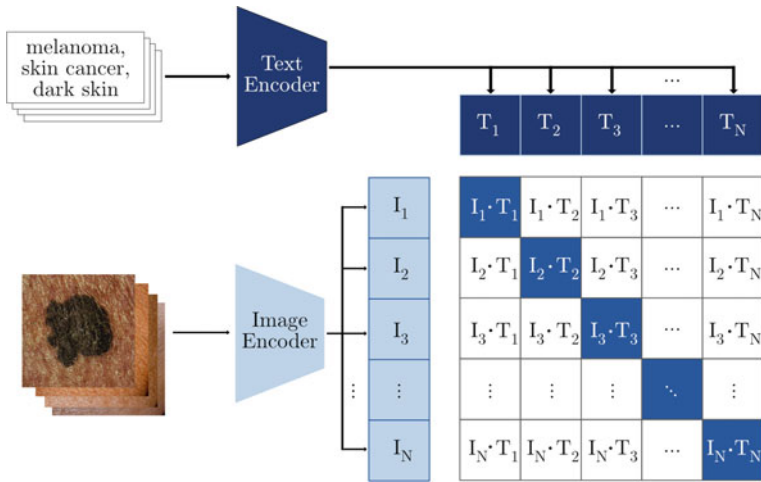
### 3.2 Zero-Shot Text-to-Image Generative Models

The main limitation of GANs is that they can only produce data for a known domain. Although the models achieve realistic results when applied in denoising [43], chest X-ray generation [44], CT [45], or MRI [37], the use of random variable noise does not replace the closed-world assumption of the training dataset [46]. Besides, there are domain transfer approaches for applying few-shot GAN models [47], but these are not widely adopted and are out of the scope of this study. For all of them we can mention Choi et al. [48] and Li et al. [49].

Zero-data learning was established in Larochelle et al. [50] as an attempt to address a prediction of samples from classes, which were not observed during training. This can be achieved by passing additional information that encodes properties of objects associating observed and non-observed classes [51]. One of modern models solving this problem is CLIP (Contrastive Language-Image Pre-training), model trained on

---

<sup>1</sup> <https://www.isic-archive.com/>.



**Fig. 3** Simultaneous training of the text and image encoder to learn visual representations from natural language supervision

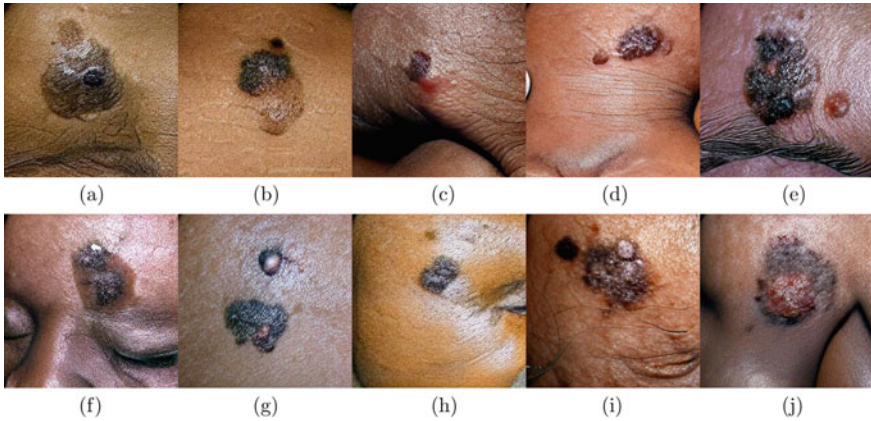
the images and their text descriptions pairs from internet [52]. The model learns through encoders the best abstract representation of images and text and their combination, as shown in Fig. 3.

**DALL-E 2** DALL-E 2, introduced in Ramesh et al. [53], is an AI model from OpenAI, the successor of DALL-E presented in Ramesh et al. [54], that can create realistic images and art from a description provided in English. The newer version reduces number of parameters from 12 billion parameters to 3.5 billion and generate even more realistic images with higher resolution and enable a new functionality called inpainting: generating patterns according to text input directly into a provided image [55]. DALL-E 2 is achieving an excellent performance thanks to the combination of diffusion models [56], CLIP image embeddings, CLIP text embeddings, and GPT-3 (Generative Pre-trained Transformer 3), the language model introduced in Brown et al. [57]. GPT architecture is based on the transformer model type, which uses the attention technique [58]. In an effort to prevent misuse of DALL-E, OpenAI excluded sexual and violent content from the training set and blocks prompts with their explicit mention.<sup>2</sup>

## 4 Experiments and Results

Zero-Shot Text-to-Image autoregressive models were used to generate synthetic medical data of patients with Fitzpatrick skin types IV–VI, as shown in Fig. 4. For

<sup>2</sup> <https://github.com/openai/dalle-2-preview/blob/main/system-card.md>.



**Fig. 4** Artificially generated data showing melanoma in Fitzpatrick skin type IV–VI created using Zero-Shot Text-to-Image model

the experiment, 10 generated images of melanoma were randomly selected. These images were assessed by an expert dermatologist to determine whether they show characteristics used for the diagnosis of melanoma.

#### **4.1 Dermatological Perspective**

The ABCD rule, introduced in 1985, is one of the most common methods used to identify potentially malignant melanoma [59]. The acronym stands for Asymmetry (two halves don't match), Border (borderline irregularity), Color (changes in color) and Diameter (often larger than 6 mm). The ABCD acronym was expanded in 2004 to include the letter E, which stands for Evolving [60]. As shown on Fig. 1, each criterion has certain characteristics that are monitored to distinguish between benign and malignant lesion. In addition, this technique failed to detect several malignant nevi in their earlier stages [19, 21].

The evaluation of a skin lesion from a single image, at a given quality, is difficult. In clinical evaluation, it is common practice to investigate the lesion during personal examination using dermatoscope, which allows better observation of edges, internal structure, colors, angiogenesis, etc. [61].

For each of the generated 10 melanoma images for patients with Fitzpatrick skin type IV–VI, the possible type of melanoma and the characteristics indicative of it (with the exception of rule **D**: Diameter, which is difficult to determine from the single image) were assessed by an experienced dermatologist. Table 2 lists possible differential diagnoses that may be benign as well as malignant. As part of the assessment, the examiner was only provided with the visual data shown in Fig. 4.

**Table 2** Assessment of 10 selected generated melanoma images for Fitzpatrick skin type IV–VI by an experience dermatologist

	(a)	(b)	(c)	(d)	(e)
dg.	Superficial spreading malignant melanoma/initial nodular melanoma	Superficial spreading malignant melanoma	Superficial spreading malignant melanoma	Superficial spreading malignant melanoma, nodular melanoma with satellite metastasis	Superficial spreading malignant melanoma, nodular melanoma with satellite metastasis
diff. dg.	Dysplastic nevus, seborrheic keratosis	Dysplastic nevus	Pigmented junctional melanocytic nevus, dysplastic nevus	Seborrheic keratosis, pigmented basal cell carcinoma	Seborrheic keratosis, pigmented basal cell carcinoma
A	Irregular shape	Irregular shape	Irregular shape	Irregular shape	Irregular shape
B	Irregular border	Irregular border	Irregular border	Irregular border	Irregular border
C	Multicolored or uneven coloring	Multicolored or uneven coloring	Multicolored or uneven coloring	Multicolored or uneven coloring	Multicolored or uneven coloring
D	(Size cannot be evaluated)	(Size cannot be evaluated)	(Size cannot be evaluated)	(Size cannot be evaluated)	Large in size
	(f)	(g)	(h)	(i)	(j)
dg.	Lentigo maligna melanoma, superficial spreading malignant melanoma, initial nodular melanoma	Superficial spreading malignant melanoma, nodular melanoma with satellite metastasis	Superficial spreading malignant melanoma	Nodular melanoma with satellite metastasis	Nodular melanoma
diff. dg.	Seborrheic keratosis	Seborrheic keratosis with keratine pearls, dysplastic nevus	Pigmented junctional melanocytic nevus, dysplastic nevus	Seborrheic keratosis with melanocytic nevus, pigmented basal cell carcinoma with melanocytic nevus	Seborrheic keratosis, pigmented basal cell carcinoma, squamous cell carcinoma
A	Irregular shape	Irregular shape	–	Irregular shape	Irregular shape
B	Irregular border	Irregular border	–	Irregular border	Irregular border
C	Multicolored or uneven coloring	Multicolored or uneven coloring	Dark brown/black colour	Multicolored or uneven coloring	Multicolored or uneven coloring
D	Large in size	(Size cannot be evaluated)	(Size cannot be evaluated)	(Size cannot be evaluated)	(Size cannot be evaluated)

The assessor determined the possible type of melanoma and relevant characteristics important for the evaluation. The table also provides possible differential diagnoses



**Table 3** Distribution of individual lesions across the training dataset for the ResNet50V2 multi-class classifier

Class	AKIEC	BCC	BKL	DF	MEL	NV	VASC	Total
n	332	514	1099	115	1563	3061	142	6826

## 4.2 CAD Perspective

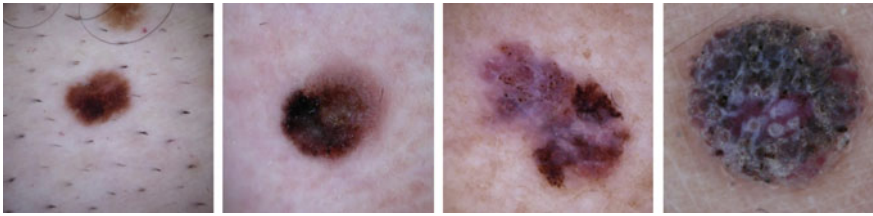
Residual networks (ResNets) are unique type of deep convolutional networks whose basic idea is to skip blocks of convolutional layers by using shortcut connections [62]. In this case study, we use a variant of the residual neural networks called ResNet50V2 [63]. ResNets provide tradeoff between performance and number of parameters. The weights used in the proposed model have been pre-trained using the ImageNet database [63].

The ResNet architecture for computer vision applications is based on two simple rules: (i) the layers share the same number of filters for the same output feature map size; and (ii) the number of filters is doubled when the feature map size is halved [62]. The down-sampling is performed by convolutional layers that perform a stride of 2 and batch normalization is carried right after each convolution operation and before ReLU activation [62]. The identity shortcut is used when the input and output are of the same dimensions. The projection shortcut is used to match dimensions through  $1 \times 1$  convolutions when the dimensions increase. When the shortcuts go across feature maps of two sizes, they are performed with a stride of 2 [62]. The network uses a fully-connected layer and softmax function at the output [63].

**Training Dataset** The training dataset shown in Table 3 consists of 6826 dermatoscopic images from the Medical University of Vienna dataset (HAM10000) [64]. The dataset represents a significant part of important diagnostic categories: actinic keratosis, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanoma in various stages (Fig. 5), melanocytic nevus, and vascular lesions (hemorrhages, pyogenic granulomas, angiokeratomas or angiomas).

In a significant proportion of the images ( $\sim 50\%$ ), the ground truth was defined by histopathological analysis of the tissue, in the remaining images the class was determined either by expert consensus or by *in vivo* confocal microscopy. Dermatologic images were divided between the training and validation sets in an 80/20 ratio.

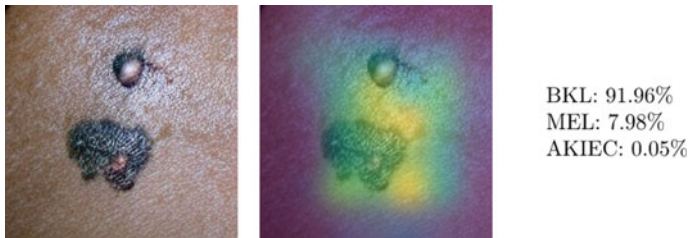
**Results** Despite the limited knowledge of the target domain, the used model classified 7 synthetic images of Fitzpatrick skin type IV–VI patients as melanoma and 2 images as high-risk melanocytic nevus, shown in Table 4. In only one case (image g) the image model predicted the skin lesion as benign keratosis with high confidence probability. The heatmap localization shown in Fig. 6 suggests that the classifier might face issues with more findings or their spatial distribution is irregular. The prediction results together with high confidence may suggest that the synthetically generated data possess some of the characteristics typical for diagnosis of melanoma.



**Fig. 5** Examples of melanoma at different stages represented in the training set

**Table 4** Results of ResNet50V2 model prediction on synthetic melanoma images in patients with Fitzpatrick skin type IV–VI

Image	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)
Model prediction	MEL	MEL	NV	NV	MEL	MEL	BKL	MEL	MEL	MEL
Confidence (%)	99.98	100	99.99	100	100	100	91.96	83.92	100	99.98



**Fig. 6** Incorrect prediction (image g), heatmap localization and the top 3 predictions of the ResNet50V2 multi-class classifier

## 5 Conclusions

Although automatic detection and evaluation of skin findings is the subject of active research, the limited availability of datasets containing images of patients with Fitzpatrick skin types IV–VI for training clinically robust machine learning-driven models has not received much attention.

In our study, we explored the use of Zero-Shot Text-to-Image generative models to produce synthetic data for more accurate and fair diagnosis of melanoma in patients with Fitzpatrick skin types IV–VI. Our experiment was subsequently validated by (a) an experienced dermatologist, and (b) a CAD system using the ResNet50V2 multi-class classification model trained on the preselected HAM10000 dataset.

Synthetic data may represent a reliable solution for augmenting existing datasets to improve the performance of other AI tools aimed at improving early diagnosis of skin cancer in underrepresented populations. The approach could be applied to patients with Fitzpatrick skin type IV–VI, where data collection is more difficult due to more limited cases.

**Conflict of interest** In relation to this study, we declare the following conflicts of interest: the research was funded by Carebot s.r.o.

## References

1. Lerner, B., Stewart, L., Horowitz, D. & Carvajal, R. Mucosal Melanoma: new insights and therapeutic options for a unique and aggressive disease. *Oncology (08909091)*. **31** (2017)
2. D'Arcy, C., Holman, J. & Armstrong, B. Pigmentary traits, ethnic origin, benign nevi, and family history as risk factors for cutaneous malignant melanoma. *Journal Of The National Cancer Institute*. **72**, 257–266 (1984)
3. McGovern, V., Mihm Jr, M., Bailly, C., Booth, J., Clark Jr, W., Cochran, A., Hardy, E., Hicks, J., Levene, A., Lewis, M. & Others The classification of malignant melanoma and its histologic reporting. *Cancer*. **32**, 1446–1457 (1973)
4. Rastrelli, M., Tropea, S., Rossi, C. & Alaibac, M. Melanoma: epidemiology, risk factors, pathogenesis, diagnosis and classification. *In Vivo*. **28**, 1005–1011 (2014)
5. Lopes, F., Sleiman, M., Sebastian, K., Bogucka, R., Jacobs, E. & Adamson, A. UV exposure and the risk of cutaneous melanoma in skin of color: a systematic review. *JAMA Dermatology*. **157**, 213–219 (2021)
6. Gohara, M. Skin cancer in skins of color: Skin cancer. *Journal Of Drugs In Dermatology*. **7**, 441–445 (2008)
7. Wu, X., Eide, M., King, J., Saraiya, M., Huang, Y., Wiggins, C., Barnholtz-Sloan, J., Martin, N., Cokkinides, V., Miller, J. & Others. Racial and ethnic variations in incidence and survival of cutaneous melanoma in the United States, 1999–2006. *Journal Of The American Academy Of Dermatology*. **65**, S26–e1 (2011)
8. Gupta, A., Bharadwaj, M. & Mehrotra, R. Skin cancer concerns in people of color: risk factors and prevention. *Asian Pacific Journal Of Cancer Prevention: APJCP*. **17**, 5257 (2016)
9. Rigel, D., Friedman, R. & Kopf, A. The incidence of malignant melanoma in the United States: issues as we approach the 21st century. *Journal Of The American Academy Of Dermatology*. **34**, 839–847 (1996)
10. Coit, D., Andtbacka, R., Bichakjian, C., Dilawari, R., DiMaio, D., Guild, V., Halpern, A., Hodi, F., Kashani-Sabet, M., Lange, J. & Others Melanoma. *Journal Of The National Comprehensive Cancer Network*. **7**, 250–275 (2009)
11. Hida, T., Kamiya, T., Kawakami, A., Ogino, J., Sohma, H., Uhara, H. & Jimbow, K. Elucidation of melanogenesis cascade for identifying pathophysiology and therapeutic approach of pigmentary disorders and melanoma. *International Journal Of Molecular Sciences*. **21**, 6129 (2020)
12. Bastian, B. The molecular pathology of melanoma: an integrated taxonomy of melanocytic neoplasia. *Annual Review Of Pathology: Mechanisms Of Disease*. **9** pp. 239–271 (2014)
13. Apalla, Z., Nashed, D., Weller, R. & Castellsagué, X. Skin cancer: epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches. *Dermatology And Therapy*. **7**, 5–19 (2017)
14. Shortliffe, E. & Sepúlveda, M. Clinical decision support in the era of artificial intelligence. *Jama*. **320**, 2199–2200 (2018)
15. Jain, A., Way, D., Gupta, V., Gao, Y., Oliveira Marinho, G., Hartford, J., Sayres, R., Kanada, K., Eng, C., Nagpal, K. & Others Development and assessment of an artificial intelligence-based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. *JAMA Network Open*. **4**, e217249–e217249 (2021)
16. Liu, Y., Jain, A., Eng, C., Way, D., Lee, K., Bui, P., Kanada, K., Oliveira Marinho, G., Gallegos, J., Gabriele, S. & Others A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*. **26**, 900–908 (2020)

17. Nahar, V., Ford, M., Jacks, S., Thielen, S., Johnson, A., Brodell, R. & Bass, M. Sun-related behaviors among individuals previously diagnosed with non-melanoma skin cancer. *Indian Journal Of Dermatology, Venereology And Leprology*. **81** pp. 568 (2015)
18. Sober, A. & Burstein, J. Precursors to skin cancer. *Cancer*. **75**, 645–650 (1995)
19. Carli, P., Massi, D., Giorgi, V. & Giannotti, B. Clinically and dermoscopically featureless melanoma: when prevention fails. *Journal Of The American Academy Of Dermatology*. **46**, 957–959 (2002)
20. Li, Y., Esteva, A., Kuprel, B., Novoa, R., Ko, J. & Thrun, S. Skin cancer detection and tracking using data synthesis and deep learning. ArXiv Preprint [ArXiv:1612.01074](https://arxiv.org/abs/1612.01074). (2016)
21. Liu, W., Hill, D., Gibbs, A., Tempany, M., Howe, C., Borland, R., Morand, M. & Kelly, J. What features do patients notice that help to distinguish between benign pigmented lesions and melanomas?: the ABCD (E) rule versus the seven-point checklist. *Melanoma Research*. **15**, 549–554 (2005)
22. Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H. & Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. **542**, 115–118 (2017)
23. Mar, V. & Soyer, H. Artificial intelligence for melanoma diagnosis: how can we deliver on the promise?. *Annals Of Oncology*. **29**, 1625–1628 (2018)
24. Sun, M., Kentley, J., Mehta, P., Dusza, S., Halpern, A. & Rotemberg, V. Accuracy of commercially available smartphone applications for the detection of melanoma. *British Journal Of Dermatology*. **186**, 744–746 (2022)
25. Combalia, M., Codella, N., Rotemberg, V., Carrera, C., Dusza, S., Gutman, D., Helba, B., Kittler, H., Kurtansky, N., Liopyris, K. & Others Validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: the 2019 International Skin Imaging Collaboration Grand Challenge. *The Lancet Digital Health*. **4**, e330–e339 (2022)
26. Haenssle, H., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A., Thomas, L., Enk, A. & Others Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals Of Oncology*. **29**, 1836–1842 (2018)
27. Masood, A., Al-Jumaily, A. & Anam, K. Self-supervised learning model for skin cancer diagnosis. *2015 7th International IEEE/EMBS Conference On Neural Engineering (NER)*. pp. 1012–1015 (2015)
28. Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., Gallivanone, F., Cozzi, A., D’Amico, N. & Sardanelli, F. AI applications to medical images: From machine learning to deep learning. *Physica Medica*. **83** pp. 9–24 (2021)
29. Ghassemi, M., Oakden-Rayner, L. & Beam, A. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*. **3**, e745–e750 (2021)
30. FDA Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). (Department of Health, 2019)
31. Tjoa, E. & Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions On Neural Networks And Learning Systems*. **32**, 4793–4813 (2020)
32. Handa, A., Patraucean, V., Badrinarayanan, V., Stent, S. & Cipolla, R. Understanding real world indoor scenes with synthetic data. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 4077–4085 (2016)
33. Jaderberg, M., Simonyan, K., Vedaldi, A. & Zisserman, A. Synthetic data and artificial neural networks for natural scene text recognition. ArXiv Preprint [ArXiv:1406.2227](https://arxiv.org/abs/1406.2227). (2014)
34. Benaim, A., Almog, R., Gorelik, Y., Hochberg, I., Nassar, L., Mashichi, T., Khamaisi, M., Lurie, Y., Azzam, Z., Khoury, J. & Others Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. *JMIR Medical Informatics*. **8**, e16492 (2020)
35. El Emam, K., Mosquera, L., Fang, X., El-Hussuna, A. & Others Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study. *JMIR Medical Informatics*. **10**, e35734 (2022)

36. Shin, H., Tenenholz, N., Rogers, J., Schwarz, C., Senjem, M., Gunter, J., Andriole, K. & Michalski, M. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. *International Workshop On Simulation And Synthesis In Medical Imaging*. pp. 1–11 (2018)
37. Peng, Y., Chen, S., Qin, A., Chen, M., Gao, X., Liu, Y., Miao, J., Gu, H., Zhao, C., Deng, X. & Others Magnetic resonance-based synthetic computed tomography images generated using generative adversarial networks for nasopharyngeal carcinoma radiotherapy treatment planning. *Radiotherapy And Oncology*. **150** pp. 217–224 (2020)
38. Onishi, Y., Teramoto, A., Tsujimoto, M., Tsukamoto, T., Saito, K., Toyama, H., Imaizumi, K. & Fujita, H. Automated pulmonary nodule classification in computed tomography images using a deep convolutional neural network trained by generative adversarial networks. *BioMed Research International*. **2019** (2019)
39. Han, X., Zhang, L., Zhou, K. & Wang, X. ProGAN: Protein solubility generative adversarial nets for data augmentation in DNN framework. *Computers & Chemical Engineering*. **131** pp. 106533 (2019)
40. Narayanan, H., Dingfelder, F., Butté, A., Lorenzen, N., Sokolov, M. & Arosio, P. Machine learning for biologics: opportunities for protein engineering, developability, and formulation. *Trends In Pharmacological Sciences*. **42**, 151–165 (2021)
41. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. Generative adversarial networks. arXiv e-prints. ArXiv Preprint [ArXiv:1406.2661](https://arxiv.org/abs/1406.2661). **1406** (2014)
42. Limeros, S., Majchrowska, S., Zoubi, M., Rosén, A., Suvilehto, J., Sjöblom, L. & Kjellberg, M. GAN-based generative modelling for dermatological applications-comparative study. ArXiv Preprint [ArXiv:2208.11702](https://arxiv.org/abs/2208.11702). (2022)
43. Chen, S., Shi, D., Sadiq, M. & Cheng, X. Image denoising with generative adversarial networks and its application to cell image enhancement. *IEEE Access*. **8** pp. 82819–82831 (2020)
44. Khalifa, N., Taha, M., Hassanien, A. & Elghamrawy, S. Detection of coronavirus (covid-19) associated pneumonia based on generative adversarial networks and a fine-tuned deep transfer learning model using chest x-ray dataset. ArXiv Preprint [ArXiv:2004.01184](https://arxiv.org/abs/2004.01184). (2020)
45. Jin, Q., Cui, H., Sun, C., Meng, Z. & Su, R. Free-form tumor synthesis in computed tomography images via richer generative adversarial network. *Knowledge-Based Systems*. **218** pp. 106753 (2021)
46. Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X. & Wang, F. Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal Of Automatica Sinica*. **4**, 588–598 (2017)
47. Kim, T., Cha, M., Kim, H., Lee, J. & Kim, J. Learning to discover cross-domain relations with generative adversarial networks. *International Conference On Machine Learning*. pp. 1857–1865 (2017)
48. Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S. & Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 8789–8797 (2018)
49. Li, X., Luo, M., Ji, S., Zhang, L. & Lu, M. Evaluating generative adversarial networks based image-level domain transfer for multi-source remote sensing image segmentation and object detection. *International Journal Of Remote Sensing*. **41**, 7343–7367 (2020)
50. Larochelle, H., Erhan, D. & Bengio, Y. Zero-Data Learning of New Tasks. *Proceedings Of The 23rd National Conference On Artificial Intelligence - Volume 2*. pp. 646–651 (2008)
51. Mensink, T., Gavves, E. & Snoek, C. Costa: Co-occurrence statistics for zero-shot classification. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 2441–2448 (2014)
52. Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. (arXiv, 2021)
53. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. (arXiv,2022)

54. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. & Sutskever, I. Zero-shot text-to-image generation. *International Conference On Machine Learning*. pp. 8821–8831 (2021)
55. Kapelyukh, I., Vosylius, V. & Johns, E. DALL-E-Bot: Introducing Web-Scale Diffusion Models to Robotics. ArXiv Preprint [ArXiv:2210.02438](https://arxiv.org/abs/2210.02438). (2022)
56. Croitoru, F., Hondru, V., Ionescu, R. & Shah, M. Diffusion models in vision: A survey. ArXiv Preprint [ArXiv:2209.04747](https://arxiv.org/abs/2209.04747). (2022)
57. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. Language Models are Few-Shot Learners. *Advances In Neural Information Processing Systems*. **33** pp. 1877–1901 (2020)
58. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances In Neural Information Processing Systems*. **30** (2017)
59. Nachbar, F., Stolz, W., Merkle, T., Cognetta, A., Vogt, T., Landthaler, M., Bilek, P., Braun-Falco, O. & Plewig, G. The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal Of The American Academy Of Dermatology*. **30**, 551–559 (1994)
60. Jensen, J. & Elewski, B. The ABCDEF rule: combining the “ABCDE rule” and the “ugly duckling sign” in an effort to improve patient self-screening examinations. *The Journal Of Clinical And Aesthetic Dermatology*. **8**, 15 (2015)
61. Wolner, Z., Yélamos, O., Liopyris, K., Rogers, T., Marchetti, M. & Marghoob, A. Enhancing skin cancer diagnosis with dermoscopy. *Dermatologic Clinics*. **35**, 417–437 (2017)
62. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 770–778 (2016)
63. Yamazaki, M., Kasagi, A., Tabuchi, A., Honda, T., Miwa, M., Fukumoto, N., Tabaru, T., Ike, A. & Nakashima, K. Yet another accelerated sgd: Resnet-50 training on imagenet in 74.7 seconds. ArXiv Preprint [ArXiv:1903.12650](https://arxiv.org/abs/1903.12650). (2019)
64. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*. **5**, 1–9 (2018)

# BD-Transformer: A Transformer-Based Approach for Bipolar Disorder Classification Using Audio



Mohamed Ramadan, Hazem Abdelkawy, Mustaqueem, and Alice Othmani

**Abstract** Bipolar disorder, named also manic depression, is a mental disorder perceived by extreme mood swings that include abnormally emotional lows (depression) and highs (hypomania or mania episodes). According to the World Health Organization, 46 million people around the world have bipolar disorder, including 2.8% of the U.S. population. The risk of suicide in bipolar disorder over a period of 20 years is high, 6% died by suicide, while 30–40% engaged in self-harm. Bipolar disorder (BD) is a significant public health issue and computer-aided diagnosis systems are needed for the diagnosis and the follow-up of patients. In this paper, a new Transformer-based approach for BD classification based on audio data is proposed. Our proposed approach outperforms all existing approaches for BD classification on the turkish audio-visual bipolar disorder corpus by achieving an accuracy of 88.2% and a F1-score of 87.8%.

**Keywords** Computer-aided diagnosis · Bipolar disorder · Audio analysis · Computer vision

## 1 Introduction

Mental Disorder is defined as disturbance in an individual's cognition, or behavior based on the world health organization (WHO) [1]. It's usually connected with stress or impairment in important areas of functioning. In 2019, one in eight individuals, or 970 million people worldwide, had a mental illness. Anxiety and depressive disorders are the most prevalent [2] and increased due to the COVID-19 pandemic in 2020. Initial projections indicate an increase of 26% in anxiety and 28% in major depressive disorders, in just one year [3]. Although there are effective methods for

---

M. Ramadan · H. Abdelkawy · A. Othmani (✉)  
Université Paris-Est Créteil (UPEC), LISSI, Vitry sur Seine 94400, France  
e-mail: [alice.othmani@u-pec.fr](mailto:alice.othmani@u-pec.fr)

Mustaqueem  
Interaction Technology Laboratory, Department of Software, Sejong University, Seoul 05006, South Korea

both prevention and therapy, the majority of those who suffer from mental illnesses do not have access to them. Stigma, prejudice, and human rights violations are also commonplace. For bipolar disorder, as a type of mental illness, it's stated that about 40 million people have its symptoms in 2019 [2]. Depressive episodes and times of manic symptoms alternate for people with bipolar illness. The person has a depressed mood (feels gloomy, irritable, or empty) or loses interest in activities for the majority of the day, almost every day, during a depressive episode. Suicide risk is higher for those who have bipolar disorder. Nevertheless, there are effective therapeutic methods available, such as psychoeducation, stress reduction and social functioning enhancement, and medication. In general, mental health is one of the most neglected areas of health globally.

Distortion of bipolar (BD) is a historically referred psychiatric disorder, called bipolar depression. In young adults, BD is the most prevalent psychological condition in the top ten of a disability-adapted life year. Consequently, it is extremely important that BD episodes can be detected early and accurately through machine learning technologies. The current methods of mental illness diagnosis are primarily based on psychological interviews and very subjective self-reported ratings. An automated recognition system helps to identify symptoms at an early stage and offers insights into biological diagnostic markers [4–10].

There is a great need to pay attention to mental diseases by the side of the scientific community. More researches are needed to be done over patients trying to help them and detect the mental disease early. All information related to patients and all details about the patients' mental health levels are needed to be captured and easily accessed. Deepening our understanding of mental health issues is made possible by artificial intelligence (AI) and machine learning (ML), which are also potential tools for supporting psychiatrists in improved clinical judgment and analysis [11]. AI approaches have demonstrated superior performance in a variety of data-rich implementation environments in recent years, including bipolar disorder [4–10, 12, 13].

In this paper, we propose a new and robust approach for bipolar disorder recognition. We developed a model using audio recordings for patients who are diagnosed with bipolar disorder clinically. We apply deep learning mechanisms on the audio data to extract useful features that helps disease recognition. We trained the model with labeled data and test it with new unseen files. We implement different approaches to reach the optimum classification accuracy. The contributions in this paper are the new audio transformer-based model, as well as the comparative experimental outcomes for various problem-solving strategies. The paper is organised as follows. Section 2 presents related works on BD diagnosis methods using machine learning. A detailed description of our proposed approach is presented in Sect. 3. Experiments and results are presented in Sect. 4. Section 5 concludes this research work.



## 2 Related Work

The bipolar disorder (BD) recognition system is a recent dynamic field of research, and researchers have designed numerous approaches over the last few years for a robust and significant system. The use of advanced deep learning approaches, researchers mostly used complex hand-crafted features for bipolar with the conventional machine learning approaches such support vector machine (SVM) and k-nearest neighbor (kNN). Therefore, researchers are greatly motivated by the growing use and the high performances of deep learning methods for bipolar disorder tasks. Hence, [14] focuses on the AVEC-18 Bipolar Disorder Challenge and analyze a number of BD Corpus modalities and introduced a new model for the hierarchical reminder for different phases where a patients with different manic levels are re-called. Furthermore, the authors [15] has suggested a multi-modal deep learning system to analyze automatically signs of mental illness utilized audiovisual and textual data. A multi-DDAE approach used for encoding per frame representations over a variety of audiovisual features as well as compact per-session descriptors using Fisher Vector Encoder. The authors improved the performance by integration of the interview transcripts provided to Paragraph Vector (PV) models as a multi-task learning system to handle overfitting. In this regards, we proposed an efficient multi-modal representation learning system for identification of bipolar disorder on the basis of depression detection to overcome the limitation without being directly optimized to the learning task [15].

Nowadays, Deep learning is certainly the most common method of speech processing but it is still not extensively used in healthcare applications. Hence, small companies with health problems are also available to provide the data, in 2018 the audiovisual challenge collect 218 audio samples of 46 individuals for Bipolar Disorder Corpus [16]. Moreover, the brain disorder in BD causes mood changes that prevent patients from doing ordinary daily tasks [17]. In this work, we overcome the problem and categorized patients suffering from BD into one of the three categories or episodes: hypomania, mania and remission using various profound strategies, feature fusion, and connecting techniques along with a basic sliding window protocol and obtained a positive results as compared [17]. Furthermore, the authors [18] focused on the audiovisual Emotion Challenge (AVEC-18) as a Bipolar Disorder Challenge (BDC) task. They proposed two new features: A histogram-dependent arouses, Long Short-Term Memory Neural Method (LSTM-RNN) measures the continuous arousal values by Deep Neural Networks and Random Forestry used for Ensemble Learning.

BD is a prevalent mental disorder that affects the job and social function in a negative way. The bipolar symptoms are episodic, particularly when there are irregular differences between episodes making the accurate diagnosis of BD very difficult. The authors [19] introduced a new audio-based technique named IncepLSTM, which incorporates the LSTM and Inception module for feature sequence to capture temporal multi-scale cues. The authors suggested a severity-sensitive loss based on a triple loss to set-up the inter-severity relationship which help to get a representative and discriminating representation of BD severity [19]. Additionally, some researchers

identify patients with bipolar disorders in remission, hypo-manic and manic audio-visual recordings of organized interviews into the 2018 audio and visual emotion recognition challenge (AVEC) [20]. In this work, the authors suggested ‘turbulence characteristics,’ to catch shifts in audio and visual contours suddenly, and to demonstrate its effectiveness for the task at hand. In [21] authors used a Capsule Neural Network (CapsNet) to identify BD patients in three categories: recovery, hypomania, and mania followed by an episode of mania. The capsNet seeks the critical spatial hierarchy between the spectrum’s of audio files, to overcome the limitations of Convolutions Neural Networks (CNN’s). These capsules attempt to represent input data meaningfully and learn the right BD class as a vector of operation that show the output of each capsule.

Furthermore, the scientists undertake the task of detecting BD states by tracking effective data derived from organized interview video recordings. Our objective is the classification of the condition of BD patients into clinically important recovery, hypomania, and mania states. In order to derive facial characteristics from video signals, the author used a Convolution Neural Network (CNN) model [22] in a hybrid mode for BDC using Visual Information. Recently, the “Bipolar Disorder and Intercultural Effect Identification” audio-visual emotional challenge and workshop (AVEC 2018) is the eighth contest for the comparison of multimedia processing and automated audio-visual well-being and emotion analytic learning methods, with participants exclusively participating under the same conditions [23]. The goal is to provide a shared test set for multi-modal data processing, to put together community health and emotional awareness as well as video processing cultures, and to compare the relative merits of different approaches to health and emotional awareness from real-life data. Three proposed tasks have been proposed in this challenge: classification of bipolar disturbances, cross-cultural emotional identification, and generation of emotional labeling from individual ratings. In this paper, we propose a novel intelligent end-to-end transformer based bipolar disorder recognition system to overcome the limitations of data scariness and optimization. In the subsequent section, we give the detailed description of our proposed approach.

### 3 Proposed Framework

In this section, the different steps of the proposed approach are detailed. First, a pre-processing of splitting into smaller chunks is applied to the audio recordings. Augmentation has been used on the audio files after they are split, in order to increase training data diversity. After pre-processing, features are extracted from the audio files to describe the key characteristics or patterns of BD. The chosen model is wav2vec2 framework for self-supervised learning of representations from raw audio data. In the subsequent sections, we extensively describe each component of our proposed approach.

### 3.1 Preprocessing

Pre-processing is a very crucial step in our proposed approach, because of the high ratio signal to noise and the limited number of training samples. We have applied pre-processing on the raw audio files represented into two major steps; splitting and augmentation. Splitting divides, the files into smaller chunks based on pre-defined files for each recording and based on the concept of avoiding training the model with very large files. Augmentation helps removing noise from the audio files and improve learning by enforcing consistency.

**Audio Splitting** Data splitting is one of the basic steps for model training and our dataset contain audio files with multiple recordings for all patients. In suggested dataset, one patient may have more than one recording. The recording length is not fixed with separator files to show the sound of a tone that is used for bipolar disorder (BD) detection. The audio length ranges from 14s up to around 17 min. We split the audio files based on those separators (tones). Also, we split the audio files to be more than 1 s and less than 60s length. A file that is less than a second in length is dropped and a file that is longer than 60s is split into several chunks as long as the chunks do not exceed 60s in length. Also, another splitting step is to use a tone in the audio files as a separator. The timing of such separators is predefined in a CSV file and used to split the files into chunks without such tone. Splitting is not windowing, it's just a step to enrich the dataset by more consistent audio files.

**Audio Augmentation** The data scarceness and quality affect the model performance. Audio augmentation is a group of techniques that create changed copies of an existing audio files and add them to it or generate new data using simulation techniques or also deep learning techniques like the generative adversarial network (GAN).

Data Augmentation in general and audio augmentation techniques more specifically generate a wide range of natural data variances and can function as a Regularizer to lessen the issue of over-fitting. Additionally, it can assist deep neural networks become resilient to intricate variances in real-world data, which enhances their generalization capabilities.

We utilized the time dropout (chunk drop), and frequency dropout (freq. drop) techniques as an augmentation to augment the dataset for training. In the time dropout method, drops chunks and replaces some random chunks of the original waveform with zeros. In the frequency dropout, it drops frequency, instead of adding zeros in the time domain and adds zeros in the frequency domain.

### 3.2 Feature Extraction

Once data is pre-processed, feature extraction step is performed to extract high level feature from input tensor. The used model is wav2vec model with previously set

weights; in other words, a pre-trained model. In fact, using pre-trained or non-pre-trained model, features do not differ as they are the same features the network would have extracted from the data and the weights represent the network understanding of the data and how it can learn from the dataset. The use of weights of a pre-trained neural network only accelerate the model training and help to overcome the problem of data scarcity.

We have chosen wav2vec model [24] to benefit from the self-supervised learning mechanism. Self-supervised learning has become a paradigm in machine learning for learning general data representations from unlabeled samples and refining the model with tagged data. We used Wav2Vec2 model as the main transformers' architecture. Wav2Vec2 is a transformer-based architecture for automatic speech recognition (ASR) tasks [24]. Using a novel contrastive pre-training objective, Wav2Vec2 learns discriminating speech representations. The model of the wav2vec2 takes the audio signal and passes it through CNN model to extract very good representative features from it. The feature extractor converts the speech signal to the model's input format, and afterwards a tokenizer converts the model's output format to text. They are both required since ASR models convert speech to text.

Features are fed into transformers neural network that can learn from unlabeled speech. The Transformer generate initial representations, or embedding, for each audio file. Next, it aggregates information from all of the other audio files, generating a new representation per file. The self-attention computation requires as input a query  $Q$ , keys  $K$ , and values  $V$ . The embedding matrices  $X$  are multiplied with the corresponding internal weight matrices  $W_Q$   $W_K$   $W_V$  as defined in Eq. 1. Let  $I \in Q, K, V$

$$I = X \times W. \quad (1)$$

The self-attention is thus described in Eq. 2:

$$attention(Q, K, V) = softmax(Q \cdot K^T)V \quad (2)$$

In fact, Eq. 2 is simply the dot-product attention function. The scaling factor is ignored for simplicity. This step is repeated several times for all files, successively generating new embeddings. Transformers here generate contextualized representations, which means they will perform a classification and a forecasting process at the same time. Then the model is fine-tuned on labeled data with the Connectionist Temporal Classification (CTC) algorithm for specific ASR tasks. CTC is a neural network that associate a scoring function, for training RNNs networks. It allows to handle sequence problems with variable sequence/time size. The workflow of our proposed Transformer-based model is summarized in Fig. 1.



**Fig. 1** System model workflow from input; training steps and generating accuracy based on test set

## 4 Experiments and Results

### 4.1 Dataset

The dataset used to evaluate the performance of the proposed approach is a Turkish dataset called Bipolar Disorder Corpus (BDC) and introduced by Ciftci et al. [25]. The Bipolar Disorder Corpus (BDC) is gathered from patients from the mental health service of a hospital. It has been annotated by psychiatrists with BD states and the Young Mania Rating Scale (YMRS) scores. Those scores were acquired at session level where each score corresponds to one patient on one of the test days.

The BDC has been used in the AVEC 2018 Challenge [23], a competition event for proposing and comparing new data processing and machine learning methods for bipolar disorder diagnosis. A subset of BDC has been used in the AVEC challenge, it contains video recordings of clinical interviews of 100 Turkish locals who have been diagnosed with a specific type of BD. We are concerned with only audio data in our experiments.

These patients were chosen from a mental health facility with the ethical committee's agreement and had a prior BD diagnosis following the DSM-5 inclusion criteria. Patients who met certain exclusion criteria, such as abusing drugs or alcohol three months previous to the onset of another severe organic disease, exhibiting hallucinogenic symptoms, having low mental capacity, or acting disruptively during the session, were excluded [25]. Additionally, several participants refused to have their information made public. The total number of subjects in the publicly available dataset is 46, with a mean age of 36.5 and a standard variation of 10.2 years. There are 30 men and 16 women in this final group. Sessions were documented both during the patient's hospital stay and after their release in the third month. After each session, a YMRS score is given to each subject where:  $YMRS \leq 7$  is Remission,  $YMRS \in [8, 19]$  is Hypomania, and  $YMRS \geq 20$  is diagnosed as Mania.

In our experiments, we used the 218 audio recordings, made publicly available in the AVEC challenge, which will be then split into training, validation and test subsets. For more details, please refer to training explanation given in the next section.

## 4.2 Implementation Details

**Train/Validation/Test split:** we assigned 80% of the data for training, 10% for validation, and 10% for model testing to show the significance and robustness of the proposed model.

**Fine-tuning strategy:** To overcome the data scarcity problem and to reduce the computational time of training, a transfer learning strategy is set up with two steps:

- pre-training step: the deep neural networks are first trained on a first and big dataset that owns enough labeled audio signals for a related task.
- fine-tuning step: the deep neural networks are initialized with the parameters learnt in the first pre-training step and then trained a second time for the new task. In this work, the CNNs are fine-tuned on the Bipolar Disorder Corpus dataset to test.

**The best hyperparameters of our Wav2Vec2 transformer:** the learning rate is heuristically tuned until fine-tuning has become stable. We set it as 0.00001 in all our trials. We use the number of training epochs of 16, 32 and 128 during our tries. Seed Values of 0, 1234, 1993 are set fix. Max Length of the recording to be processed of 10, 20, 30, 45, and 60 s are considered. A batch size of 2, 4, and 8 are evaluated in our experimentation. The tested hyperparameters are shown in Table 1.

## 4.3 Performance of the Proposed Approach for Bipolar Disorder Diagnosis

In this section, we present the experiment results of our transformer-based approach for Bipolar Disorder diagnosis. Several experiments have been performed to find the best configuration and hyperparameters to learn accurate audio patterns of bipolar disorder. Table 1 shows a comparison of the trials made with Transformers. We have used a pre-trained model and we tested also to train the model from scratch. We tested different parameters such as the number of training epochs, learning rate, batch size, and max length of audio files. The best results reached an accuracy of 88.2% in the test set.

Our transformer-based deep neural network has shown very accurate results in the training and it achieves an accuracy of 95% after 16 epochs and no over-fitting problem is reported. It has been demonstrated also that it has a very good generalization capacity in the test set by achieving a precision of 87.9%, a recall of 87.7%, a F1-score of 87.8% and an accuracy of 88.2%.

**Table 1** Experimental results for transformers

Method	Augmented use	Seed	Epochs	Batch size	Max length (s)	Accuracy (%)
wav2vec2_large	N/A	0	32	2	10	45.80
wav2vec2_large	N/A	0	32	2	20	51.10
wav2vec2_base	N/A	0	32	4	60	61.30
wav2vec2_base	N/A	0	128	4	30	65.40
wav2vec2_base	N/A	0	32	8	30	84.70
wav2vec2_base	N/A	1993	16	4	30	75.10
wav2vec2_base	N/A	1234	16	4	45	69.00
wav2vec2_base	Time dropout	1234	16	4	30	82.90
wav2vec2_base	Time and frequency dropout	1234	16	4	30	<b>88.20</b>

#### 4.4 Comparison with State of the Art Methods

We compared the performances of our Wav2Vec2-Base model and existing approaches for BD classification using audio data [17, 19, 21] on the same data (the Turkish BD corpus). Table 2 demonstrates that the pre-trained Wav2Vec2-Base achieved 0.882 overall accuracy and UAR 0.877, which is better than the state-of-the-art approaches.

Du et al. [19] utilized audio and developed model that received the lowest UARs in the AVEC 2018 Challenge [23]. For the purpose of diagnosing bipolar disorder, [21] employed audio spectrograms, and they obtained an UAR of 46.2%. Sequential models such as LSTM, RNNs, and BiLSTM architectures are trained using the baseline characteristics supplied by the competition organizers with a low-level audio features (MFCCs, eGEMAPS, Bag-of-Acoustic-Words (BoAW), and DeepSpectrum).

They achieved a 74.60% UAR on the training set when MFCC and BoVW features are fused on Bi-LSTM model. However, the stated test set result of 33.33% is the UAR score at the chance level for the 3-class classification issue. By avoiding over-training, deep learning models should be trained with caution of LSTM that is trained using all of the characteristics offered in the challenge achieves the greatest results on the test set. Success is indicated by a 59.24% UAR on the validation and a 44.44% UAR on the test under such conditions. But this result is worse than the SVM model's baseline performance on the dataset due to sequence structure and learning strategy [17].

On a limited dataset, employing more sophisticated deep learning models does not always improve performance. A contrastive task defined through quantization of latent representations is solved with wav2vec2, which masks the speech input in the latent space. Our proposed wav2vec2-transformer-based approach achieve an accuracy of 88.2% and an UAR of 87.7% and it outperforms existing methods on Turkish BD dataset despite the limited number of samples in this dataset.

**Table 2** Performance comparison of bipolar disorder classification and assessment methods on AVEC2018 (BD) dataset

Method	Metrics	Audio features	Model used	Performance (%)
[21]	UAR	Raw audio	CapsNet	45.5
[17]	UAR	Deep spectrum features	LSTM/Bi-LSTM	58.20
[19]	UAR	MFCCs coefficients	InceptLSTM/LSTM	65.1
Our	UAR	Raw audio	Transformers	87.7

## 5 Conclusion and Future Work

In this paper, a new approach for BD diagnosis and classification using audio data is proposed. After pre-processing the audio signals by splitting and augmenting the signal data, they are fed to a Wav2vec transformer based deep neural networks to learn BD patterns and to classify the audio signals into three BD classes. Thanks to the transformer's capacity to capture long-range dependencies and interactions, the proposed approach outperforms existing approaches on the Turkish audio-visual BD corpus by achieving 88.2% of accuracy. In future work, we are planning to propose a multi-modal approach for BD diagnosis that fuses audio, visual and textual patterns extracted from video data.

## References

1. <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>
2. <https://vizhub.healthdata.org/gbd-results/>
3. [https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci\\_Brief-Mental\\_health-2022.1](https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1)
4. Muzammel, M., Othmani, A., Mukherjee, H., Salam, H.: Identification of signs of depression relapse using audio-visual cues: A preliminary study. In: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), pp. 62–67. IEEE (2021)
5. Muzammel, M., Salam, H., Hoffmann, Y., Chetouani, M., Othmani, A.: Audvowelconsnet: A phoneme-level based deep cnn architecture for clinical depression diagnosis. *Machine Learning with Applications* **2**, 100,005 (2020)
6. Muzammel, M., Salam, H., Othmani, A.: End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. *Computer Methods and Programs in Biomedicine* **211**, 106,433 (2021)
7. Othmani, A., Kadoch, D., Bentounes, K., Rejaibi, E., Alfred, R., Hadid, A.: Towards robust deep neural networks for affect and depression recognition from speech. In: International Conference on Pattern Recognition, pp. 5–19. Springer (2021)
8. Othmani, A., Zeghina, A.O.: A multimodal computer-aided diagnostic system for depression relapse prediction using audiovisual cues: A proof of concept. *Healthcare Analytics* **2**, 100,090 (2022)
9. Othmani, A., Zeghina, A.O., Muzammel, M.: A model of normality inspired deep learning framework for depression relapse prediction using audiovisual data. *Computer Methods and Programs in Biomedicine* **226**, 107,132 (2022)



10. Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., Othmani, A.: Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control* **71**, 103,107 (2022)
11. Lin, E., Lin, C.H., Lane, H.Y.: Precision psychiatry applications with pharmacogenomics: Artificial intelligence and machine learning approaches. *International journal of molecular sciences* **21**(3), 969 (2020)
12. Divya, M., Ankalkoti, P.: Bipolar classification methodology deep learning
13. Fernandes, B.S., Karmakar, C., Tamouza, R., Tran, T., Yearwood, J., Hamdani, N., Laouamri, H., Richard, J.R., Yolken, R., Berk, M., et al.: Precision psychiatry with immunological and cognitive biomarkers: a multi-domain prediction for the diagnosis of bipolar disorder or schizophrenia using machine learning. *Translational psychiatry* **10**(1), 1–13 (2020)
14. Xing, X., Cai, B., Zhao, Y., Li, S., He, Z., Fan, W.: Multi-modality hierarchical recall based on gbdt for bipolar disorder classification. In: *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pp. 31–37 (2018)
15. Zhang, Z., Lin, W., Liu, M., Mahmoud, M.: Multimodal deep learning framework for mental disorder recognition. In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 344–350. IEEE (2020)
16. Ren, Z., Han, J., Cummins, N., Kong, Q., Plumbley, M.D., Schuller, B.W.: Multi-instance learning for bipolar disorder diagnosis using weakly labelled speech data. In: *Proceedings of the 9th International Conference on Digital Public Health, DPH2019*, p. 79–83. Association for Computing Machinery, New York, NY, USA (2019). DOI 10.1145/3357729.3357743. <https://doi.org/10.1145/3357729.3357743>
17. Ebrahim, M., Al-Ayyoub, M., Alsmirat, M.: Determine bipolar disorder level from patient interviews using bi-lstm and feature fusion. In: *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 182–189. IEEE (2018)
18. Yang, L., Li, Y., Chen, H., Jiang, D., Oveneke, M.C., Sahli, H.: Bipolar disorder recognition with histogram features of arousal and body gestures. In: *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pp. 15–21 (2018)
19. Du, Z., Li, W., Huang, D., Wang, Y.: Bipolar disorder recognition via multi-scale discriminative audio temporal representation. In: *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pp. 23–30 (2018)
20. Syed, Z.S., Sidorov, K., Marshall, D.: Automated screening for bipolar disorder from audio/visual modalities. In: *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pp. 39–45 (2018)
21. Amiriparian, S., Awad, A., Gerczuk, M., Stappen, L., Baird, A., Ottl, S., Schuller, B.: Audio-based recognition of bipolar disorder utilising capsule networks. In: *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE (2019)
22. Abaei, N., Al Osman, H.: A hybrid model for bipolar disorder classification from visual information. In: *ICASSP*, vol. 2020, pp. 4107–4111 (2020)
23. Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A., et al.: Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In: *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, pp. 3–13 (2018)
24. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations (2020). 10.48550/ARXIV.2006.11477. <https://arxiv.org/abs/2006.11477>
25. Çiftçi, E., Kaya, H., Güleç, H., Salah, A.A.: The Turkish audio-visual bipolar disorder corpus. In: *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pp. 1–6 (2018). <https://doi.org/10.1109/ACIIAsia.2018.8470362>

# Establishment and Analysis of a Combined Diagnostic Model of Acute Myocardial Infarction Based on Random Forests and Artificial Neural Networks



Zhenrun Zhan, Xiaodan Bi, Jinpeng Yang, Xu Tang, and Tingting Zhao

**Abstract** Acute myocardial infarction is a serious disease worldwide that kills approximately 8.5 million patients each year. It can occur in multiple age groups and, despite the more diverse diagnostic techniques available, it has a number of limitations. Therefore, a diagnostic model based on gene biomarkers should be developed to assist existing diagnostic methods and improve the efficiency of diagnosis. For this research, we applied three datasets, one for screening DEGs and the other two for validation. We selected the DEGs of AMI from the first dataset and used a random forest classifier to identify key genes, including TREM-like transcript 2 (TREM2), interleukin-1 receptor type 2, CSF3R, HMGB2, nuclear factor interleukin 3 regulated, granzyme K (GZMK), MXD1, KIAA1324, NTNG2, and LOC440737. Among these genes, TREM2, HMGB2, GZMK, MXD1, KIAA1324, NTNG2, and LOC440737 have never been associated with AMI. Next, we successfully used an artificial neural network to construct a new model to diagnose AMI and verified the diagnostic effect of the model using the two validation datasets.

**Keywords** Acute myocardial infarction · Biomarkers · Artificial neural networks · Random forests · Machine learning

## 1 Introduction

Acute myocardial infarction has been responsible for the largest number of deaths worldwide in the last decade [1]. This disease is the most grievous representation of acute coronary syndrome (ACS). Each year, over 2.4 million patients die from this disease in the US and over 4 million in Europe and Northeast Asia. Further, one-third of patients in developed countries die of AMI every year [2], and this number is still increasing annually. From a pathophysiological perspective, acute

---

Z. Zhan · X. Bi · J. Yang · X. Tang · T. Zhao (✉)  
Changzhi Medical College, Changzhi, Shanxi, China  
e-mail: 649823325@qq.com

Heping Hospital Affiliated to Changzhi Medical College, Changzhi, Shanxi, China

myocardial infarction can be classified into two types: STEMI and NSTEMI [3]. Both of them are collectively referred to as ACS with unstable angina pectoris. The pathophysiology of NSTEMI is similar to that of unstable angina pectoris. These conditions are collectively referred to as non-ST-segment elevation ACS.

In general, myocardial infarction is caused by the rupture of fragile atherosclerotic plaques or erosion of coronary endothelial cells [4]. Ultimately, cardiomyocyte necrosis remains the final result. AMI is still the most severe form of coronary heart disease. In 2010 alone, over 1.1 million patients in the USA were admitted to hospital because of AMI. The economic burden on society exceeded \$450 billion [5]. The diagnosis of myocardial infarction depends on the changes in the electrocardiography (ECG) results caused by myocardial ischemia or infarction; the biochemical indices related to myocardial infarction, as well as the occurrence of ischemic symptoms, are also important signals [6, 7]. First, ECG is one of the most frequently used techniques for diagnosing AMI, and its accuracy depends heavily on the clinical experience and machine operation of ECG doctors, which leads to subjective results. Second, CK-MB (Creatine Kinase Isoenzymes) and cTn (cardiac troponin) are the most distinguished biomarkers for the diagnosis of AMI [8, 9]. However, CK-MB seems relatively insensitive in detecting small myocardial infarctions. Consequently, a new diagnostic model needs to be established to fill the gap in existing techniques [10, 11]. Thereafter, an artificial neural network (ANN) model was established to predict the genetic diagnostic model of AMI.

## 2 Materials and Methods

### 2.1 Data Collection

We opened the website “[www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo)” to find appropriate datasets, used the GEOquery package to download the chip data of the GSE61144, GSE97320, and GSE48060 datasets, and integrated their clinical phenotypes and expression profiles (Table 1). We obtained the corresponding annotation information for the respective platform chip probes from the GEO database.

### 2.2 Differential Expression and Enrichment Analysis

Using the limma package to determine the difference between the 7 AMI samples and 10 normal samples from the GSE61144 dataset. The conditions for significance of the DEGs were set as follows: log FC values of  $> 1$  and  $P$ -values of  $< 0.05$ . The cluster profiler package from the R software was used for the GO and KEGG enrichment analysis of the DEGs.

**Table 1** Data download

Data	Sample size	Organization type	Data type
GSE 61144	24 (normal: 10; disease: 7)	ACS STEMI PCI blood: 7 ACS STEMI blood: 7 Normal control blood: 10	Microarray
GSE 97320	6 (normal: 3; disease: 3)	Peripheral blood of acute myocardial infarction: 3 Peripheral blood of healthy people: 3	Microarray
GSE 48060	52 (normal: 21; disease: 31)	Peripheral blood, patient without recurrent events: 26 Peripheral blood, patient with recurrent events: 5 Peripheral blood, normal control: 21	Microarray

### 2.3 *Random Forest Screening for the Important Genes*

The Random Forest package of R software was employed to establish a random forest model based on DEGs. We set the number of best-fit variables for the binary tree to 6 and selected 30 as the optimal number of trees to be included in the random forest. When the importance value > 0.6, the top 10 genes were selected to construct the model. Clustering of unsupervised hierarchical clusters of 10 significant genes was reclassified using the pheatmap package, and a heat map was drawn.

### 2.4 *PPI Network Analysis*

The STRING software was used to analyze the PPI network of the 10 DEGs. We logged in to the website ([www.string-db.org/](http://www.string-db.org/)), selected “multiple proteins by names/identifiers,” entered the names of the candidate genes, and selected the “organism” and “Homo sapiens” options; thereafter, “SEARCH” was clicked.

### 2.5 *Neural Network for Building the Disease Classification Model*

The GSE61144 dataset was used to build the ANN model. We used the NeuralNet package (version 1.44.2) and NeuralNetTools package (version 1.5.3) to build the ANN model predicated on the important variables. The disease classification model of AMI was established using the obtained gene weight information. The classification score was obtained using the following method: For the upregulated genes, the

score was evaluated as 1 point when the gene expression level was greater than its median value and 0 points when the gene expression level was lower than its median value, and vice versa. We then opened the pROC software package in the R software and imported the disease classification score to compute the verification of the AUC categorization properties.

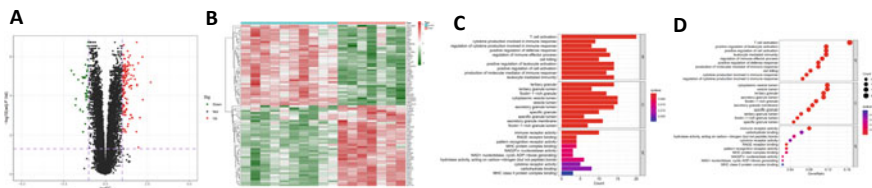
## 2.6 Additional Data Verification

The validity of the categorical score models of AMI and normals was verified using the two independent datasets (GSE97320 and GSE48060). The ROC curve of each dataset was drawn using the pROC software package, and the AUC of the two independent validation datasets was computed to test the classification efficiency.

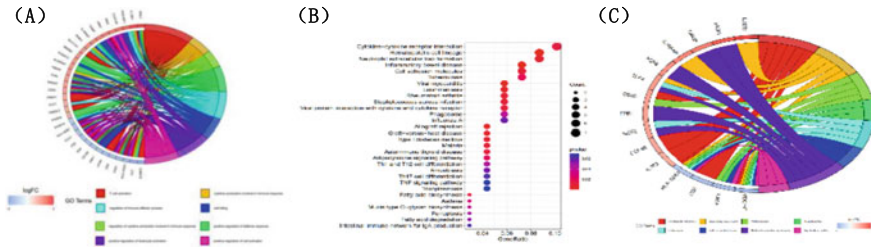
## 3 Results

### 3.1 Differential Expression Analysis

For this research, we first downloaded the chip dataset GSE61144 and analyzed the data to filter for the DEGs. The GSE61144 dataset includes 17 samples, which can be divided into 7 AMI disease samples and 10 normal samples. Thereafter, the DEGs between the AMI samples and normal samples of the chip dataset were identified using the limma software package to perform the Bayesian test. By setting the screening conditions to a significance threshold of  $P$ -values of  $< 0.05$  and log fold change (FC) values of  $> 1$ , we obtained 168 remarkable DEGs associated with AMI through conditional screening (Supplementary Document 1). The DEGs are reflected in the volcano map shown in Fig. 1a and heat map shown in Fig. 1b.



**Fig. 1** **a** Volcano plot of differential expression analysis results. **b** Heatmap of DEGs. Bar chart (c) and bubble plot (d) of GO enrichment results



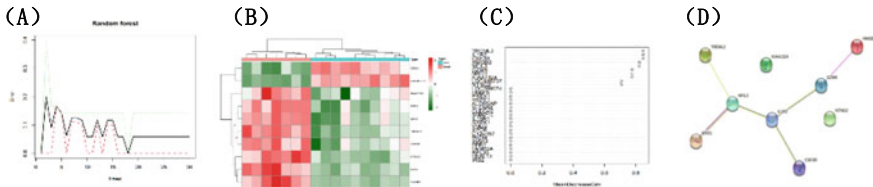
**Fig. 2** a Ring plot showing GO enrichment. The connecting line indicates that the gene is included in the GO term. Bubble chart (b) and ring plot (c) showing the KEGG pathway enrichment

### 3.2 GO/KEGG Enrichment Analysis

We used the clusterProfiler software package to perform enrichment analysis of DEGs. GO enriched the molecular functions, cellular components and biological processes, and the results are displayed in Fig. 1c, d. We found that the biological processes associated with AMI were abundant, including T cell activation, regulation of cytokine production involved in immune response, positive regulation of leukocyte activation, and leukocyte-mediated immunity cell killing. The cellular components involved included tertiary granules and other important components. The molecular functions included enrichment of immune receptor activity and other major functions. Figure 2b, c shows the results of the KEGG pathway enrichment analysis of the DEGs, including the related strikingly enriched biological pathways and relevant DEGs.

### 3.3 Random Forest Screening for DEGs

A random forest classifier was then used to process the 168 DEGs. We classified all probable numbers in 1–168 variables by cyclic random forests and computed the average error rate of this model. All variables were included to calculate the average error rate, and the results are shown in Fig. 3a. Finally, the variable coefficient was set to 6. We attempted to control the number of variables and minimize the out-of-band error as much as possible. Figure 3 (relationship plot) illustrates the relationship with the number of decision trees and the model error. To make the error in the model most stable, we set the condition to be more important than 0.6 and obtained 10 DEGs as possible genes for follow-up analysis. Figure 3c shows that TREM-like transcript 2 (TREM2), interleukin-1 receptor type 2 (IL1R2), and CSF3R were the most important variables, followed by HMGB2, nuclear factor interleukin 3 regulated (NFIL3), granzyme K (GZMK), MYC-associated factor X dimerization protein 1 (MXD1), KIAA1324, LOC440737, and NTNG2. We used these 10 significant candidates from the GSE61144 dataset to conduct K-means unsupervised clustering. Of the 17 samples in the GSE61144 dataset, the 10 genes



**Fig. 3** **a** The effect of the decision tree number on the error rate. **b** Heatmap of unsupervised clustering. **c** Findings from the random forest classifier in the Gini coefficient method. **d** Diagram of PPI network analysis of DEGs

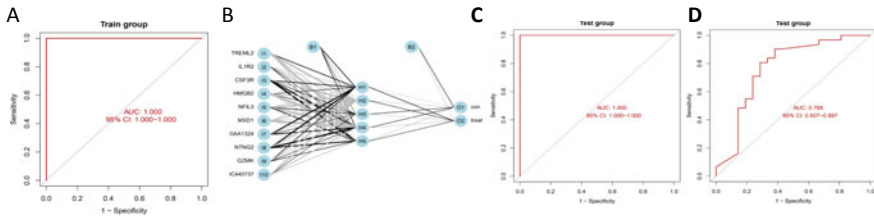
may be employed for distinguishing disease samples from normal samples. Interestingly, TREML2, IL1R2, HMGB2, NFIL3, MXD1, KIAA1324, NTNG2, and CSF3R were significantly more expressed in disease samples than in normal samples. The opposite was true for LOC440737 and GZMK.

### 3.4 Protein–Protein Interaction (PPI) Network Analysis

The STRING software was used to analyze the PPI network of the 10 DEGs. PPI refers to the non-covalent binding of mannoproteins with two or more protein molecules. We logged in to the website ([www.string-db.org/](http://www.string-db.org/)), selected “multiple proteins by names/identifiers,” entered the names of the candidate genes, and selected the “organism” and “Homo sapiens” options; thereafter, “SEARCH” was clicked. Nine key proteins were identified in the STRING software, with NFIL3 and IL1R2 being the most important proteins (Fig. 3d).

### 3.5 Construction of the ANN Model

We then used the NeuralNet software package to establish the ANN model based on GSE61144. Initially, we preprocessed the data to standardize them. Before training the neural network, we applied the min-max method [0,1] to separate the zoom data. We standardized the minimum and maximum data values, and then performed the operation. To assess the results of the neural network model more effectively, we applied statistical methods for verification. Finally, the output of the first hidden layer in the neural network results (i.e., input of the last output layer) was regarded as the result of the gene weight. Based on such, we established a model to sort the expression data of genes between AMI and normal samples. The ROC curve of the validation results shows the performance of the classification model in Fig. 4a. The AUC of the verification results was 1, demonstrating the model’s robustness. Thus, these reliable data could be used to build a neural network model and obtain the outcomes (Fig. 4b).



**Fig. 4** **a** Validation of ROC curve of classification efficiency. **b** Neural network model. **c** AUC verification results in the GSE97320 dataset. **d** AUC verification results in the GSE48060dataset

### 3.6 Evaluation of the AUC

After the maximum and minimum standardizations of the GSE97320 and GSE48060 datasets, we used these data to verify the designed ANN model, calculate the neural AMI score, evaluate its classification efficiency, and compare the AUC. The neural AMI score was obtained using the following method. For the upregulated genes, the score was evaluated as 1 point when the gene expression level was greater than its median value and 0 points when the gene expression level was lower than its median value. For the downregulated genes, the score was evaluated as 1 point when the gene expression level was lower than its median value and 0 points when the gene expression level was greater than its median value. Figure 4 shows the ROC curve of the two independent validation datasets. The scores of the two datasets were then compared. In the GSE97320 dataset (Fig. 4c), the AUC of the neural AMI score remained 1.00; the specificity was 100%, and the sensitivity was 100%. In the GSE48060 dataset (Fig. 4d), the AUC of the neural AMI score was 0.768. Although the sensitivity and specificity were lower than those in the other dataset, they remained high. This indicates that the ANN model maintained a high robustness in both the independent validation datasets.

## 4 Discussion

In this study, we identified the DEGs associated with AMI from the GSE61144 dataset, established a random forest tree model, and used a random forest classifier to obtain 10 important DEGs. The established neural network model was used to ensure the weight of the candidate genes, and a scoring model for neural AMI, which was associated with AMI, was established. Finally, we used two independent sample datasets to determine the classification efficiency of the scoring model established. The AUC values indicated that the classification efficiency of the two datasets was significantly high, and the neural AMI model in the GSE97320 dataset had a better classification efficiency than that in the GSE48060 dataset. More interestingly, LOC440737 was not found in either verification dataset; therefore, we can conclude that LOC440737 is not a key gene for diagnosis.



Of nine genes, CSF3R encodes granulocyte colony-stimulating factor (G-CSF). After consulting the literature, we found that the receptor for this factor is G-CSFR. As a multifunctional cytokine, it functions in mutual regulation with receptors. In cardiovascular disease, this factor can promote the division of granulocytes and strengthen the function of mature neutrophils, leading to the intensification of the inflammatory response, promoting the development of the disease. In contrast, Takano et al. showed that G-CSF and its receptor can enhance the activity of bone marrow stem cells, leading to vascular and myocardial regenerations, which are conducive to the recovery of cardiac function after AMI and can reduce the associated death rate [12]. As for other aspects, CSF3R also plays a role in chronic neutrophilic leukemia, chemoresistance of hepatoblastoma, and atypical CML [13–15].

According to study, IL1R2 is a key mediator related to a variety of immune cells. The IL1R2 gene can act as a mediator of cell metabolism and many other immune responses mediated by IL1R2, and it also affects atherosclerosis [16–18]. In another study, IL1R2 expression increased in cells with myocardial ischemia-reperfusion injury. IL1R2 also inhibits the function of IL-17RA; this mechanism reduces cardiomyocyte apoptosis. Therefore, overexpression of cytokine receptors encoded by IL1R2 in cardiomyocytes can serve as a basis for a new approach for the treatment and alleviation of myocardial ischemia-reperfusion injury [19].

In terms of heart disease, some authors [20–22] have conducted a preliminary study on the diagnostic biomarkers of AMI and found gene modules related to AMI. It also affects cardiac function by regulating the signal transduction process of cellular molecules and causes heart failure. This gene may affect the pathogenesis and development of heart failure through these processes and has the potential to become a new site for intervention in patients with heart failure [23]. After consulting several relevant studies [24–26], we concluded that NFIL3 can regulate the inflammatory response causing gout; thus, it is possible to treat gout through this gene.

The TREML2 gene, a triggering receptor expressed on myeloid cells, has been confirmed to be associated with the pathogenesis of Alzheimer's disease [27]. HMGB2 encodes high-mobility group box 2, the expression of which decreases with age. Jeong et al. confirmed that the protein encoded by the gene is involved in a variety of biological processes, including regulation of cell aging [28].

After reading many related articles [29–31], we obtained the following conclusions. GZMK mainly affects immune aging in the entire body by promoting the production of inflammatory cytokines. Currently, there is no research on its relationship with heart diseases. Appropriate expression of netrin-G2, which is encoded by NTNG2, is conducive to the normal development of human nerves.

Our study did not aim to design a new diagnostic model and completely replace the existing diagnostic methods, but to help improve the efficiency of such current diagnostic models. Our diagnostic model has high accuracy and sensitivity, especially in East Asians. However, previous studies have been based on STEMI datasets, and whether the diagnostic model established can be applied to the diagnosis of NSTEMI is unclear. In addition, the accuracy of the model needs to be further investigated in consideration of our results. Indeed, there are still some deficiencies in our diagnostic model, and the accuracy and specificity need to be further improved.

**Acknowledgement** This study was supported by Shanxi Province Graduate Education Innovation Project (2022Y37), and Provincial Science and Technology Grant of Shanxi Province (20210302124588).

## References

1. G. W. Reed, J. E. Rossi, and C. P. Cannon, "Acute myocardial infarction," *The Lancet*, vol. 389, no. 10065, pp. 197–210, 2017.
2. R. W. Yeh, S. Sidney, M. Chandra, M. Sorel, J. V. Selby, and A. S. Go, "Population trends in the incidence and outcomes of acute myocardial infarction," *New England Journal of Medicine*, vol. 362, no. 23, pp. 2155–2165, 2010.
3. Thygesen K, Alpert JS, Jaffe AS, Simoons ML, Chaitman BR, White HD (2012) Third universal definition of myocardial infarction. *J Am Coll Cardiol* 60(16):1581–1598
4. P. Libby, "Mechanisms of acute coronary syndromes and their implications for therapy," *N Engl J Med*, vol. 368, pp. 2004–2013, 2013.
5. W. S. Weintraub, S. R. Daniels, L. E. Burke, B. A. Franklin, D. C. Goff Jr, L. L. Hayman, D. Lloyd-Jones, D. K. Pandey, E. J. Sanchez, A. P. Schram et al., "Value of primordial and primary prevention for cardiovascular disease: a policy statement from the American heart association," *Circulation*, vol. 124, no. 8, pp. 967–990, 2011.
6. Thygesen K, Alpert JS, Jaffe AS, Simoons ML, Chaitman BR, White HD (2012) Third universal definition of myocardial infarction. *J Am Coll Cardiol*
7. Stone NJ, Robinson JG, Lichtenstein AH, Bairey Merz CN, Blum CB, Eckel RH, Goldberg AC, Gordon D, Levy D, Lloyd-Jones DM et al (2014) 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American college of cardiology/American heart association task force on practice guidelines. *J Am Coll Cardiol* 63(25 Part B):2889–2934
8. Roffi M, Patrono C, Collet J-P, Mueller C, Valgimigli M, Andreotti F, Bax JJ, Borger MA, Brotons C, Chew DP, Gencer B, Hasenfuss G, Kjeldsen K, Lancellotti P, Landmesser U, Mehilli J, Mukherjee D, Storey RF, Windecker S, ESC Group (2016) 2015 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: task force for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation of the European Society of Cardiology (ESC). *Eur Heart J* 37(3):267–315
9. H. Jneid, J. L. Anderson, R. S. Wright, C. D. Adams, C. R. Bridges, D. E. Casey, S. M. Ettinger, F. M. Fesmire, T. G. Ganiats, A. M. Lincoff et al., "2012 accf/aha focused update of the guideline for the management of patients with unstable angina/non–st-elevation myocardial infarction (updating the 2007 guideline and replacing the 2011 focused update) a report of the American college of cardiology foundation/American heart association task force on practice guidelines," *Journal of the American College of Cardiology*, vol. 60, no. 7, pp. 645–681, 2012.
10. Liu, S., Xiao, Z., You, X. and Su, R., 2022. Multistrategy boosted multicolony whale virtual parallel optimization approaches. *Knowledge-Based Systems*, 242, p.108341.
11. Su R, Gu Q, Wen T (2014) Optimization of high-speed train control strategy for traction energy saving using an improved genetic algorithm. *J Appl Math*
12. H. Takano, M. Ohtsuka, H. Akazawa, H. Toko, M. Harada, H. Hasegawa, T. Nagai, and I. Komuro, "Pleiotropic effects of cytokines on acute myocardial infarction: G-csf as a novel therapy for acute myocardial infarction," *Current pharmaceutical design*, vol. 9, no. 14, pp. 1121–1127, 2003.
13. R. Beekman and I. P. Touw, "G-csf and its receptor in myeloid malignancy," *Blood, The Journal of the American Society of Hematology*, vol. 115, no. 25, pp. 5131–5136, 2010.

14. J. E. Maxson, J. Gotlib, D. A. Pollyea, A. G. Fleischman, A. Agarwal, C. A. Eide, D. Bottomly, B. Wilmot, S. K. McWeeney, C. E. Tognon et al., "Oncogenic *csf3r* mutations in chronic neutrophilic leukemia and atypical cml," *New England Journal of Medicine*, vol. 368, no. 19, pp. 1781–1790, 2013.
15. S. Fujiyoshi, S. Honda, M. Minato, M. Ara, H. Suzuki, E. Hiyama, and A. Taketomi, "Hypermethylation of *csf3r* is a novel cisplatin resistance marker and predictor of response to postoperative chemotherapy in hepatoblastoma," *Hepatology Research*, vol. 50, no. 5, pp. 598–606, 2020.
16. A. Smith, L. Keen, M. Billingham, M. Perry, C. Elson, J. Kirwan, J. Sims, M. Doherty, T. Specter, and J. Bidwell, "Extended haplotypes and linkage disequilibrium in the *il1r1-il1a-il1b-il1rn* gene cluster: association with knee osteoarthritis," *Genes & Immunity*, vol. 5, no. 6, pp. 451–460, 2004.
17. C. A. Dinarello, "The interleukin-1 family: 10 years of discovery 1," *The FASEB Journal*, vol. 8, no. 15, pp. 1314–1325, 1994.
18. V. A. Peters, J. J. Joesting, and G. G. Freund, "IL-1 receptor 2 (*il-1r2*) and its role in immune regulation," *Brain, behavior, and immunity*, vol. 32, pp. 1–8, 2013.
19. J. Lin, Q. Li, T. Jin, J. Wang, Y. Gong, Q. Lv, M. Wang, J. Chen, M. Shang, Y. Zhao et al., "Cardiomyocyte *il-1r2* protects heart from ischemia/reperfusion injury by attenuating *il-17*-mediated cardiomyocyte apoptosis," *Cell Death & Disease*, vol. 13, no. 1, pp. 1–12, 2022.
20. J. Chen, L. Yu, S. Zhang, and X. Chen, "Network analysis-based approach for exploring the potential diagnostic biomarkers of acute myocardial infarction," *Frontiers in physiology*, vol. 7, p. 615, 2016.
21. A. P. Ambrosy, G. C. Fonarow, J. Butler, O. Chioncel, S. J. Greene, M. Vaduganathan, S. Nodari, C. S. Lam, N. Sato, A. N. Shah et al., "The global health and economic burden of hospitalizations for heart failure: lessons learned from hospitalized heart failure registries," *Journal of the American College of Cardiology*, vol. 63, no. 12, pp. 1123–1133, 2014.
22. Y.-J. Weng, D. J.-Y. Hsieh, W.-W. Kuo, T.-Y. Lai, H.-H. Hsu, C.-H. Tsai, F.-J. Tsai, D.-Y. Lin, J. A. Lin, C.-Y. Huang et al., "E4bp4 is a cardiac survival factor and essential for embryonic heart development," *Molecular and cellular biochemistry*, vol. 340, no. 1, pp. 187–194, 2010.
23. Velmurugan BK, Chang R-L, Marthandam Asokan S, Chang C-F, Day C-H, Lin Y-M, Lin Y-C, Kuo W-W, Huang C-Y (2018) A minireview of E4BP4/NFIL3 in heart failure. *J Cell Physiol* 233(11):8458–8466
24. Y. Wang, Z. Kuang, X. Yu, K. A. Ruhn, M. Kubo, and L. V. Hooper, "The intestinal microbiota regulates body composition through *nfil3* and the circadian clock," *Science*, vol. 357, no. 6354, pp. 912–916, 2017.
25. Tang H, Tan C, Cao X, Liu Y, Zhao H, Liu Y, Zhao Y (2021) NFIL3 facilitates neutrophil autophagy, neutrophil extracellular trap formation and inflammation during gout via REDD1-dependent mTOR inactivation. *Front Med* 8
26. G. Kang, H.-S. Han, and S.-H. Koo, "Nfil3 is a negative regulator of hepatic gluconeogenesis," *Metabolism*, vol. 77, pp. 13–22, 2017.
27. S.-Y. Wang, P.-Y. Gong, Y.-D. Zhang, T. Jiang et al., "The role of *trem2* in Alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 76, no. 3, pp. 799–806, 2020.
28. H.-R. Jo and J.-H. Jeong, "MicroRNA-mediated down regulation of *hmgb2* contributes to cellular senescence in microvascular endothelial cells," *Cells*, vol. 11, no. 3, p. 584, 2022.
29. L. T. Joeckel, R. Wallich, P. Martin, D. Sanchez-Martinez, F. C. Weber, S. F. Martin, C. Borner, J. Pardo, C. Froelich, and M. M. Simon, "Mouse granzyme k has pro-inflammatory potential," *Cell Death & Differentiation*, vol. 18, no. 7, pp. 1112–1119, 2011.
30. L. T. Joeckel, C. C. Allison, M. Pellegrini, C. H. Bird, and P. I. Bird, "Granzyme k-deficient mice show no evidence of impaired antiviral immunity," *Immunology and Cell Biology*, vol. 95, no. 8, pp. 676–683, 2017.
31. C. M. Dias, J. Punetha, C. Zheng, N. Mazaheri, A. Rad, S. Efthymiou, A. Petersen, M. Dehghani, D. Pehlivan, J. N. Partlow et al., "Homozygous missense variants in *ntng2*, encoding a presynaptic netrin-g2 adhesion protein, lead to a distinct neurodevelopmental disorder," *The American Journal of Human Genetics*, vol. 105, no. 5, pp. 1048–1056, 2019.

# Striped-Cross Attention Network with Implicit Semantic Knowledge for Antibody Structure Prediction



Miao Gu  and Min Liu

**Abstract** The structure of an antibody directly determines its ability to bind with the target proteins specifically. It is of great significance to obtain accurate structure data for the antibody research in the field of disease diagnosis and therapy. Since antibody structure resolution experiments are time-consuming and costly, a growing number of methods for protein structure prediction are proposed to address this problem. Although considerable improvements have been made in the general protein structure prediction, deep learning-based approaches still fail to yield sufficiently accurate antibody structures to provide functional insights and design assistance. In this paper, we firstly exploit the unique semantic information implied from the comparison between the antibody structures and the general protein structures. We further propose a Striped-Cross Attention Network (SCA-Net) to efficiently fuse the features of different local functional regions and that of global regions on the antibody. Concretely, for each target amino acid position, SCA-Net collects its contextual information preferentially at the same functional domain (CDR or FR region) of the same peptide chain (heavy or light chain). Subsequently, the contextual information is fused with the global information. We utilize 2D ResNet with dilated convolution as the backbone network for feature extraction, and then construct a parallelly structured attention network using the proposed SCA-Net component as a classifier. Finally, an end-to-end multi-task learning framework is yielded to predict the antibody structures, which are described by the inter-residue distances and orientations. Quantitative experiments on two independent antibody structure test sets suggest that the proposed method achieves the expected results.

**Keywords** Antibody structure prediction · Striped-cross attention network · Multi-task learning · Inter-residue distances and orientations

---

M. Gu · M. Liu (✉)

Department of Automation, Tsinghua University, Beijing 10084, People's Republic of China  
e-mail: [lium@tsinghua.edu.cn](mailto:lium@tsinghua.edu.cn)

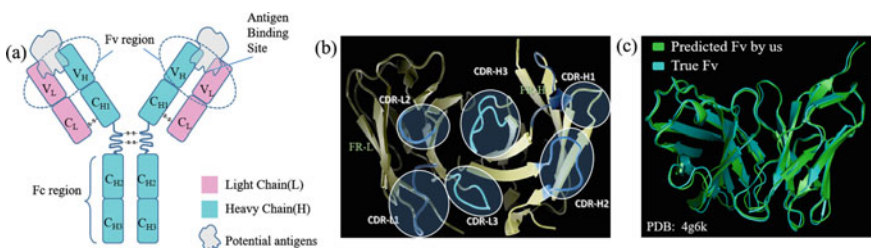
# 1 Introduction

Antibodies are a class of proteins that can specifically bind to specific target proteins and play an important role in the diagnosis and treatment of diseases today [1]. The specific binding ability of antibodies is determined by their own protein tertiary structures, and the rational design of antibodies for a given antigen relies on an accurate model of the antibody structure. As it is time-consuming and costly to perform structural analysis experiments on a large number of candidate antibody sequences during antibody design and development, antibody structure prediction has become an important research direction in protein structure prediction [2].

As shown in Fig. 1a, the antibody is a Y-shaped structural protein complex that assembled from two heavy and two light chains. The Fc region is highly conserved, it's engaged in immune effector functions, but not in antigen recognition. However, as shown in Fig. 1b, for the variable fragment (Fv) region, there exist six sequentially non-adjacent but structurally adjacent complementary decision regions (CDRs) that engaged in antigen recognition. For the reasons mentioned above, the antibody structure prediction task mainly refers to the prediction of the Fv structure.

Most of the current available methods for antibody structure prediction, including the RosettaAntibody and ABodyBuilder [3, 4], are derived from the threading methods used in common protein structure prediction, which uses resolved antibody fragment structures as homologous templates for some forms of grafting. This approach can produce models with an overall root-mean-square deviation (RMSD) less than 1 Å from the native structure [5]. However, amino acids of the CDRs is more highly variable than FRs, the available fragment template data can barely cover the conformational space of CDRs.

Recently, deep learning has made breakthroughs in the key challenging problems in structural biology including protein folding. Yang et al. [6] obtained inter-amino acid distance and angle data representation of protein structures based on protein sequence co-evolution features with deep residual network prediction. DeepMind proposed AlphaFold2 [7], an accurate prediction framework for common protein folding, based on graph network, transformer variant, and multiple sequence comparison features. David Baker et al. [8] incorporated 1D sequence level, 2D



**Fig. 1** The diagram of antibody structure, the visualization functional regions of Fv and the resulting antibody Fv structure by this work

amino acid distance map level, and 3D coordinate level information to achieve the RoseTTAFold framework, which is second only to AlphaFold2 in terms of protein structure prediction accuracy. Recently, deep learning-based methods for antibody structure prediction have been proposed [9, 10]. Borrowing the idea of ordinary protein prediction, these studies mainly devoted to the prediction of the distance and the angle between amino acids of antibodies, which are further employed as a geometric potential constraint to obtain antibody structures. These methods, to a certain degree, improve the prediction accuracy of CDR regions, while their attention models are directly derived from the image segmentation task, which do not consider the semantic-specific information in the antibody sequences.

It can be found that antibody sequences have unique semantic information compared with ordinary protein sequences. In specific, there are two types of functional regions on antibody Fv sequences, each with six fragments, which are discontinuous in sequence level but adjacent in structure. In this study, by introducing the semantic information of antibody sequences, we propose a new attention mechanism Striped-Cross Attention (SCA) for antibody structure prediction. The SCA-based module first performs attention-based feature fusion on the CDRs of antibody VH, FR of VH, CDRs of VL, and FR of VH, respectively, and then performs feature fusion between different functional domains to obtain global representations. We propose a classifier that introduces semantic information specific to antibody sequences based on the SCA module, and build a feature extraction module with Dilated Residual Network (DRN), and then implement a multi-task deep learning model for the prediction of antibody structure geometry information. After obtaining the description of geometric information, the antibody structure, as shown in Fig. 1c, can be predicted according to the pipeline of previous work [6, 9].

The main contributions of our work are as follows. (1) We proposed a new attention mechanism (SCA) for antibody structure prediction. Unlike previous attention mechanisms that perform self-attention on all amino acids equally, SCA collects contextual information for each amino acid at the same structural domain (heavy or light Chain) and the same functional domain (CDR or FR region). (2) We designed a parallel network for multiple inter-residue distances and orientations classification tasks based on the SCA module and 2D convolutional networks. (3) We combine the SCA module-based classifier with a feature extraction backbone network to implement a multi-task learning framework for the prediction task of geometric information of antibody structures. We conducted validation experiments on two test sets with highly diverse antibody structures and the experimental results showed that the proposed SCA-Net achieved encouraging results.

## 2 Methodology

In this section, the motivation of the proposed SCA-network and the constructed overall framework are briefly described. The SCA module calculates the attention

implied between amino acid pairs of the same functional domain based on the functional domain information implied by the antibody sequence compared to the normal protein sequence rather than at the global level. The overall network architecture consists of three parts: an encoder and feature splicing layer, a feature extraction backbone network, and a multi-task classifier based on the SCA module. We used this network framework to perform the task of predicting the distance and angle between antibody amino acids based on the SABDab dataset [13]. Then in line with other related work [9], antibody structures were obtained by pyRosetta based on inter-residue distances and orientations.

### 2.1 The Diagram of Striped-Cross Attention

Deep learning-based protein/antibody structure prediction tasks all rely on the amino acid pair feature matrix obtained by amino acid sequence expansion and transposition, and as mentioned previously, most remarkable protein structure prediction algorithms employ the attention mechanism to extract amino acid remote correlations. As shown in Fig. 2a, the non-local network perceives the correlation between the features at a certain position and all other positions on the feature map, and then generates the global attention-aware context information. Recently, in order to reduce the practical and spatial complexity of the attention network on the prediction of antibody structures with fewer samples, the Criss-cross attention (CCA) block

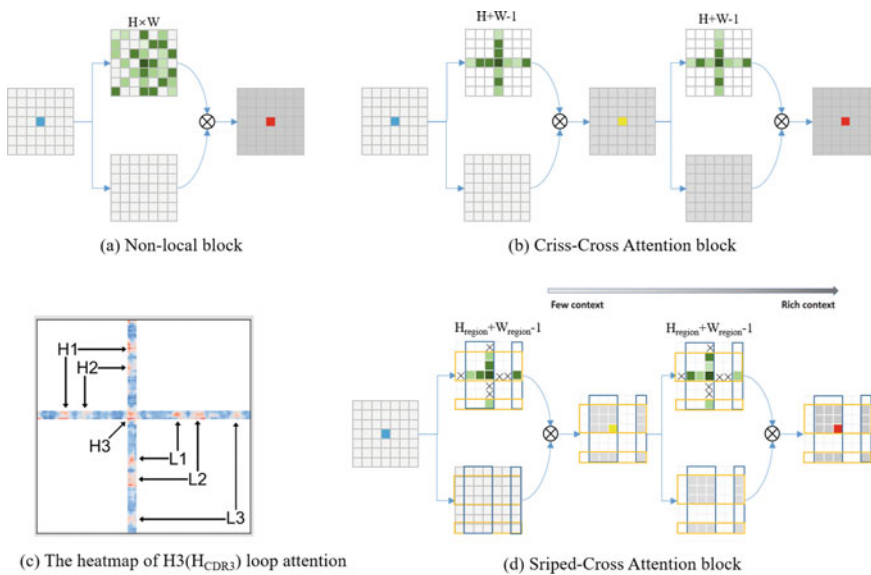


Fig. 2 Diagrams of related attention mechanism and the novel one proposed by this paper

[11], which belongs to the local-block network in the field of image segmentation, is introduced, as shown in Fig. 2b, with continuous of longitudinal and cross attention modules to collect the correlation information of each pixel point with any pixel point globally.

To some extent, the antibody sequence semantic information implies the relationship between the local structures of antibodies. The CDR regions/FR regions that are not adjacent to the sequence are structurally adjacent. Some clues about it can also be found on the attention heat map in previous studies [9], shown as Fig. 2c, where the positional attention score between CDR regions is higher than between CDR regions and FR regions. Therefore, we propose the striped-cross attention (SCA) module to embed this prior knowledge, diagrammed in Fig. 2d. For each SCA module, the horizontal and vertical stripe (yellow and purple boxes in the figure) is generated based on the functional region location information in the antibody sequence, and then the horizontal and vertical attentions are calculated at the striped-cross level for each feature point belonging to the striped-cross region.

We compare the differences between the three attentional mechanisms in Fig. 2. Specifically, all three use the input feature map with space size  $H \times W$  to generate the attention map (upper branch) and the transformed implicit feature map (lower branch), respectively and then use weighted summation to collect contextual information. The differences are: the local module adopts dense connectivity for each location at the global level for feature relevance capture, and the weight size of the attention map is  $H \times W$ . The CCA module adopts sparse connectivity for each location at the horizontal and vertical levels for feature relevance capture, and the predicted attention value matrix size is  $H + W - 1$ . Then the global two-by-two location dependency type calculation is realized after a single loop structure. The SSA module generates a series of strips for limiting the range of the sensory field based on the semantic information of the functional domain category of the antibody sequence. For each position located in the intersecting region of the strips, feature correlation is captured at the horizontal and vertical levels in the intersecting region of the strips. The predicted attention map weight size is  $H_{\text{region}} + W_{\text{region}} - 1$ , and similarly, after a cyclic structure to capture feature correlations on similar functional domains.

## 2.2 Classifier Network Structure Based on Striped-Cross Attention

Based on the SCA attention mechanism, we can perform attention feature fusion that is discontinuous at the spatial level but continuous at the functional domain level for different semantic regions of antibodies, and then fuse them with global features. In this paper, inter-amino acid distance and angle prediction classifiers based on the SCA module are constructed, and the network framework is shown in Fig. 3a. There are convolution layers for feature transformation, serial SCA modules for each function region of the antibody, and a  $1 \times 1$  convolution layer for classification.



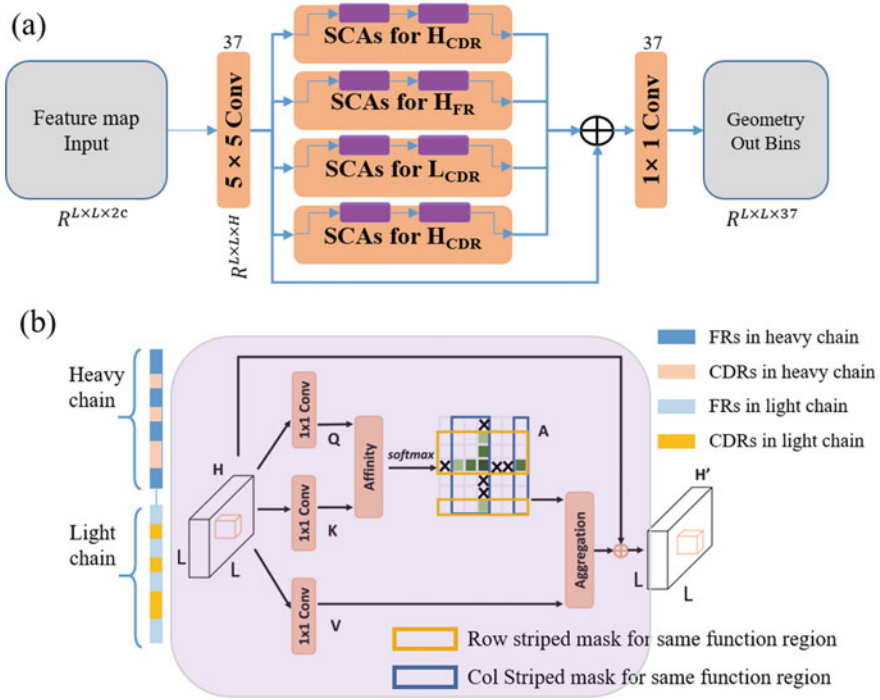


Fig. 3 Classifier network structure based on striped-cross attention

The network structure implemented for the SCA module is given in Fig. 3b. Given the amino acid pair feature map and transposed to obtain  $P_{ab} \in R^{H \times L \times L}$ , two convolutional layers with  $1 \times 1$  filters are first applied for dimensionality reduction to obtain three feature maps Q, K, and V, where  $\{Q, K\} \in R^{H' \times L \times L}$ ,  $V \in R^{H \times L \times L}$ ,  $H'$  is the number of channels whose value is smaller than H. In addition, we distinguish the antibody sequence  $\{r_0, r_1 \dots r_{L-1}\}$  into four functional domains according to light or heavy chains, CDR regions or FR regions:  $S_{Hcdr} = \{r_{0_{n_*}}^{Hcdr}, r_{1_{n_*}}^{Hcdr} \dots r_{n_*-1}^{Hcdr}\}$ ,  $S_{Hfr} = \{r_{0_{n_*}}^{Hfr}, r_{1_{n_*}}^{Hfr} \dots r_{n_*-1}^{Hfr}\}$ ,  $S_{Lcdr} = \{r_{0_{n_*}}^{Lcdr}, r_{1_{n_*}}^{Lcdr} \dots r_{n_*-1}^{Lcdr}\}$ ,  $S_{Lfr} = \{r_{0_{n_*}}^{Lfr}, r_{1_{n_*}}^{Lfr} \dots r_{n_*-1}^{Lfr}\}$ , where  $n_*$  is each index of residue at the function region. We will perform the attention calculation process on the respective scope of the four defined functional domains.

After obtaining the feature maps Q and K and the functional domain range, we further obtain the attention maps by the Striped Affinity operation to generate the respective attention maps,  $A^{region} \in R^{(n_*+n_*-1) \times n_* \times n_*}$  based on the functional domain range. The Striped Affinity operation is performed as follows: for each position u at feature maps, we obtain its feature vector  $q_u \in R^{H'}$ , and at the same time, we obtain the set  $\Omega_u \in R^{(n_*+n_*-1) \times H'}$ , combined by feature vectors located in the same row or column as that position in the given functional domain range,  $\Omega_{i,u}$  denotes the i-th element in  $\Omega_u$ . Then, for position u, we define the following striped affinity

operation:

$$d_{i,u} = q_u \Omega_{i,u}^T \tag{1}$$

$d_{i,u} \in \mathbb{R}^{(n_s+n_s-1) \times H'}$  can be used to measure the correlation degree of other location features in the functional domain with the current location  $u$ . Traversing  $i$  and  $u$  at each location of  $S_*$  and we can get the mat  $D \in \mathbb{R}^{(n_s+n_s-1) \times H'}$ , and then we calculate the local functional domain attention map  $A^{region}$  through the softmax layer.

At the same time, for position  $u$  we can obtain the feature vector  $v_u \in \mathbb{R}^H$  on the feature map  $V$ , and obtain the set of feature vectors  $\Phi_u \in \mathbb{R}^{(n_s+n_s-1) \times H}$  on the functional domain range that belongs to the same row or column as position  $u$ . Thus, the striped contextual information fusion is achieved, and this operation is defined as the Aggregation operation:

$$P_{u'} = \sum_{i \in |\Phi_u|} A_{i,u}^{region} \Phi_{i,u} + P_u \tag{2}$$

### 2.3 Framework Architecture for Antibody Structure Prediction

The overall network framework for antibody structure prediction proposed in this study is shown in Fig. 4. There are three main components: an encoder and feature splicing layer, a feature extraction backbone network, and a multi-task classifier based on the Striped-Cross Attention network.

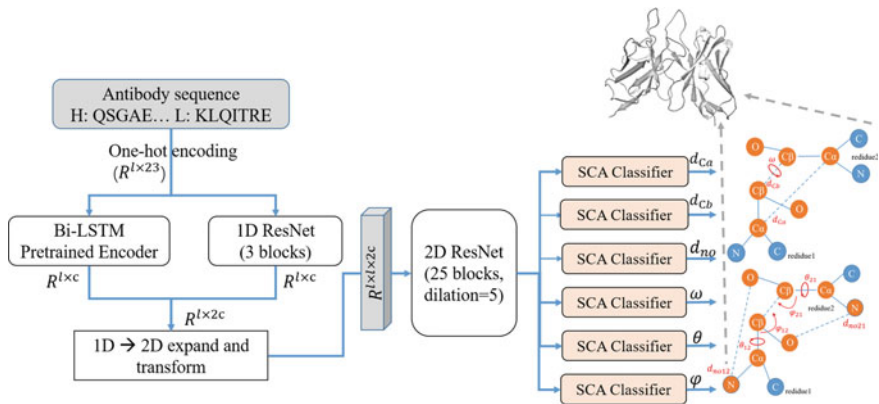


Fig. 4 The overview of our proposed network architecture

The antibody sequence is encoded in one-hot form, and the embedding representation  $p_{ab} = \{r_0, r_1 \dots r_i \dots r_{L-1}\}$ ,  $p_{ab} \in R^{L \times c}$ , is obtained by pre-trained BiLSTM [12] sequence encoder and one-dimensional convolutional neural network, where  $r_i$  represents the feature vector corresponding to the  $i$ th amino acid,  $l$  is the length of the antibody sequence, and  $c$  is the feature dimension of the hidden layer of the encoder. Then the feature vectors are tiled and longitudinally tiled to obtain the amino acid pair feature map  $P_{ab} \in R^{L \times L \times 2c}$ .

Further, through the feature extraction backbone network, we obtain the feature map  $P'_{ab} \in R^{L \times L \times 2H}$ , which implies amino acid pair interrelationship. Then the classifier based on striped-cross attention constructed in 2.2 is introduced, and the feature map  $P'_{ab}$  is fused with functional region features guided by the semantic information of functional domains on antibody sequences. Through the above classifier, we harvest 6 classes of discrete representation (37 bins) of the inter-amino acid geometric relationship matrix  $\{M_{Ca}, M_{Cb}, M_{NO}, M_{\omega}, M_{\theta}, M_{\varphi}\} \in R^{L \times L \times 37}$ . To handle the problem of the unbalanced distribution of inter-amino acid geometric relationship label values, we choose the Focal Loss as the loss function for the model training. Then, similar to the pipeline of other reports, the antibody structure was obtained by pyRosetta energy minimization based on inter-residue distance and angle information.

## 3 Experiment

### 3.1 Dataset

The model proposed in this paper was trained on the antibody database SAbDab [13] subjected to a redundancy operation (sequence identity threshold of 99%), and the database size is 1692. Two independent test sets were selected to evaluate our approach: the RosettaAntibody benchmark set (45 targets) [12] and the antibody development process work in clinical therapeutic antibodies (45 targets) [14]. Overall, the antibody assemblies in the test sets cover as much structural diversity as possible, contain therapeutic and some diagnostic antibodies, and can be used to evaluate predictive methods for antibody Fv structures.

### 3.2 Evaluation of Inter-Residue Distances and Orientations Prediction

The output of the model is the distance and orientation information between six amino acids as the geometric description of the antibody structure, and the true value of each geometric feature is discretized as 37 bins. Because amino acids that are far away from each other do not interact with each other. We adopt 18.5 Å as the

cut-off threshold for distance, i.e., all three distances of amino acid pairs larger than 18.5 Å with their corresponding angle classes are classified as the 37th bin. Besides, the outputs of the three angle matrices are positionally masked during training and validation, focusing only on the distance threshold between amino acid pairs within the relative orientation.

For this multiple output biased multi-classification problem, we calculate the average meaningful accuracy, recall, and F1 score under multi-classification for each geometry information.

$$\text{Precision}_{geo\_i} = \frac{1}{K_{geo\_i}} \sum_{j=1}^{K_{geo\_i}} \frac{TP_j}{TP_j + FP_j} \quad (3)$$

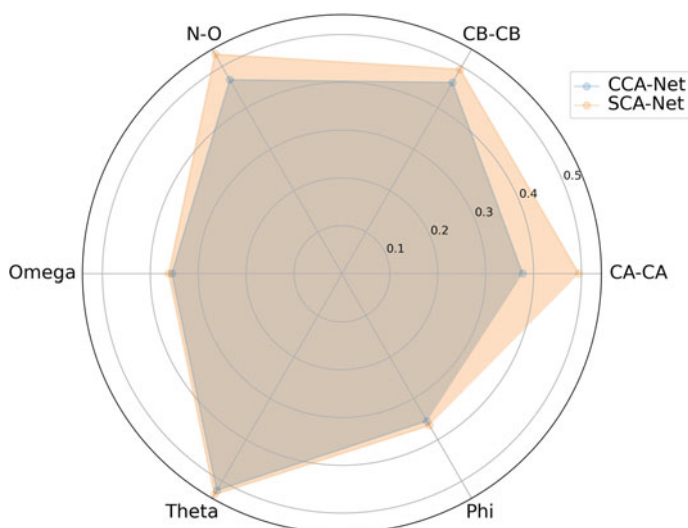
$$\text{Recall}_{geo\_i} = \frac{1}{K_{geo\_i}} \sum_{j=1}^{K_{geo\_i}} \frac{TP_j}{TP_j + FN_j} \quad (4)$$

$$\text{F1}_{geo\_i} = \frac{2 \times \text{Precision}_{geo\_i} \times \text{Recall}_{geo\_i}}{\text{Precision}_{geo\_i} + \text{Recall}_{geo\_i}} \quad (5)$$

where  $geo\_i$  represents the  $i$ th geometric information category, and  $K_{geo\_i}$  represents the number of categories of the  $i$ th information category. For the  $i$ th geometric information category,  $TP_j$  is the number of true positives for the  $j$ th category (i.e., correct prediction for that category),  $FP_j$  is the number of false positives for the  $j$ th category (i.e., misclassification of samples from other categories as samples from that category), and  $FN_j$  is the number of false negatives for the  $j$ th category (i.e., missed prediction for samples from that category). To evaluate the effectiveness of the proposed SCA, an attention mechanism guided by semantic information of antibody functional regions, we use the CCA attention mechanism to replace the SCA attention mechanism in the structure prediction framework proposed in this paper and perform comparison tests with the same hyperparameters and F1 score for classification performance evaluation, and the results are shown in Fig. 5. It can be found that the SCA attention module achieves better results in predicting geometric information between almost all amino acids, especially in distance prediction.

### 3.3 Evaluation of the Antibody Structure Prediction

We input the distance and orientation information between the six amino acids predicted by the model into pyRosetta to obtain antibody structures and compared them with those obtained by currently available antibody prediction methods. We calculated the Root Mean Square Deviation (RMSD) for the CDR loop of both chains and the backbone heavy atoms in the framework region to assess the accuracy of each functional domain, and the relative orientation coordinate distance (OCD)



**Fig. 5** Comparison of SCA-net and CCA-net at the classification task by F1 score

[5] to measure the accuracy of the relative orientation between the light and heavy chains. The evaluation results of each method on the two test sets are summarized in Table 1.

In Table 1, RosettaAntibodyG [3] and ABodyBuilder [4] are fragment-template based grafting methods, and CCA-Net (marked with an asterisk), which is the DeepAb framework proposed in related work [9], belongs to deep learning-based antibody structure prediction methods. In this work, DeepAb and the SCA-Net based

**Table 1** Performance of antibody Fv structure prediction methods on benchmarks

Method	OCD	H <sub>FR</sub> (Å)	H1 (Å)	H2 (Å)	H3 (Å)	L <sub>FR</sub> (Å)	L1 (Å)	L2 (Å)	L3 (Å)
<i>RosettaAntibody benchmark</i>									
RosettaAntibodyG	5.2	0.56	1.18	1.11	3.45	0.57	0.76	0.86	1.04
ABodyBuilder	4.72	0.49	0.96	<b>0.84</b>	<b>2.88</b>	0.49	<b>0.70</b>	0.51	1.10
CCA-Net <sup>a</sup>	5.75	0.67	1.19	1.16	3.26	0.64	1.06	0.67	1.27
Our SCA-Net <sup>a</sup>	<b>3.54</b>	<b>0.46</b>	<b>0.77</b>	0.91	2.93	<b>0.45</b>	<b>0.70</b>	<b>0.48</b>	<b>0.97</b>
<i>Therapeutic benchmark</i>									
RosettaAntibodyG	5.32	0.63	1.43	1.05	3.86	0.55	0.89	0.83	1.52
ABodyBuilder	4.25	0.49	1.02	1.01	3.09	0.45	1.05	0.51	1.35
CCA-Net <sup>a</sup>	5.26	0.61	1.23	1.10	3.31	0.57	1.17	0.65	<b>1.12</b>
Our SCA-Net <sup>a</sup>	<b>3.36</b>	<b>0.42</b>	<b>0.85</b>	<b>0.68</b>	<b>2.85</b>	<b>0.39</b>	<b>0.77</b>	<b>0.47</b>	1.17

<sup>a</sup> “Å” is a unit of distance, also written Ångström, 1Å equals 0.1 nanometers, it is a common unit to measure the accuracy of a protein’s predicted structure.

architecture proposed in this paper were trained and tested under the same hyperparameter and initialization method. The results show that our antibody structure prediction method achieves better or near-optimal results in all metrics compared to the grafting method. Moreover, it can be observed that some improvements are also achieved compared to existing deep learning antibody structure prediction frameworks, which is also consistent with the results in 3.2.

## 4 Conclusion

In this thesis, we analyze the limitations of current antibody structure prediction tasks and propose a new attention mechanism Striped-Cross Attention (SCA) embedded with semantic information of antibody sequences to efficiently fuse correlations between amino acid pair features of non-adjacent but homogeneous functional domains of antibody sequences. We constructed an antibody structure prediction framework based on this module. After validation on two independent benchmark datasets, the results show that the proposed SCA-network-based antibody prediction framework achieves more accurate results in the antibody inter-amino acid distance and angle classification tasks. The evaluation of antibody structure prediction also indicates that this method outperforms current antibody structure prediction methods. In the future, we will explore the antibody-antigen interaction prediction task based on the proposed model and graph neural networks for the case of unknown antibody structures.

**Acknowledgements** This research is supported by the National Natural Science Foundation of China (Grant No. 62173204).

## References

1. Lu, R. M., Hwang, Y. C., Liu, I. J., Lee, C. C., Tsai, H. Z., et al.: Development of therapeutic antibodies for the treatment of diseases. *Journal of Biomedical Science* 27, 1(2020).
2. Chiu, M. L., Goulet, D. R., Teplyakov, A. & Gilliland, G. L.: Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies* 8, 4(2019).
3. Sircar, A., Kim, E. T. & Gray, J. J.: RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Research* 37, W474–W479 (2009).
4. Leem, J., Dunbar, J., Georges, G., Shi, J. & Deane, C. M.: ABodyBuilder: Automated antibody structure prediction with data-driven accuracy estimation. *mAbs* 8, 7, 1259–1268 (2016).
5. Marze, N. A., Lyskov, S. & Gray, J. J.: Improved prediction of antibody VL–VH orientation. *Protein Engineering, Design and Selection* 29, 10, 409–418 (2016).
6. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., et al.: Improved protein structure prediction using predicted interresidue orientations. *PNAS-Proceedings of the National Academy of Sciences* 117, 3, 1496–1503 (2020).
7. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., et al.: Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873, 583–+ (2021).

8. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., et al.: Accurate prediction of protein structures and interactions using a three-track neural network. 373, 6557, 871–876 (2021).
9. Ruffolo, J. A., Sulam, J. & Gray, J. J.: Antibody structure prediction using interpretable deep learning. *Patterns* 3, 2(2022).
10. Mason, D. M., Friedensohn, S., Weber, C. R., Jordi, C., Wagner, B., et al. (2021) Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Zenodo*.
11. Huang, Z. L., Wang, X. G., Huang, L. C., Huang, C., Wei, Y. C., et al.: CCNet: Criss-Cross Attention for Semantic Segmentation. In: *IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 603–612). Seoul, SOUTH KOREA (2019).
12. Ruffolo, J. A., Guerra, C., Mahajan, S. P., Sulam, J. & Gray, J. J.: Geometric potentials from deep learning improve prediction of CDR H3 loop structures. *Bioinformatics* 36, 268–275 (2020).
13. Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., et al.: SAbDab: the structural antibody database. *Nucleic Acids Research* 42, D1, D1140–D1146 (2014).
14. Raybould, M. I. J., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., et al.: Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences of the United States of America* 116, 10, 4025–4030 (2019).

# A Mobile Monitoring Application for Post-traumatic Stress Disorder



Sirine Chaari, Chaima El Ouni, and Alice Othmani

**Abstract** Post-traumatic Stress Disorder (PTSD) is a impairing condition that can have a important impact on a person's life. The ability of the patient to cope with stressful situations is often affected, making PTSD treatment and monitoring indispensable. Therefore, developing a mobile application that can assist patients in identifying their symptoms and giving accurate readings is mandatory. In this paper, we propose a mental health surveillance application that will allow users to monitor their PTSD symptoms using self-assessment questionnaire and video-based technology. This application keeps track of the behavior of the patient throughout the results of the PTSD Checklist of the DSM-5 called PCL-5 test on one hand, and records videos of the patient for future artificial intelligence-based diagnosis. The web application will be publicly released.

**Keywords** Mental health · PTSD · Mobile app · PCL-5 · Patient monitoring

## 1 Introduction

Posttraumatic stress disorder (PTSD) is a mental disorder resulting after exposure to a traumatic event. These events generally involve direct threat, or represent a real risk of death or serious injury, and can cause significant distress such as natural disasters, sexual assault, military combat experience or even during a major stressful life experience like divorce or unemployment. About 4 of every 100 men develop PTSD in their lives, according to the US Department of veterans affairs. While, it is more frequent for women and it is about 8 of every 100 women, which makes it one of the most common mental health concerns worldwide. While the symptoms of PTSD may vary from one person to another, it is common for people with this disorder to have flashbacks of their traumatic event, feel and re-experience the emotions from

---

S. Chaari · C. El Ouni · A. Othmani (✉)  
LISSI, Université Paris-Est Créteil (UPEC), Vitry sur Seine 94400, France  
e-mail: [alice.othmani@u-pec.fr](mailto:alice.othmani@u-pec.fr)



the event periodically and have distressing physical reactions such as hypervigilance and self-medication with alcohol or drugs.

One of the most common tests to monitor PTSD symptoms is the Post-traumatic Stress Disorder Checklist (PCL)-5 test. It is a 20-item, widely used DSM-correspondent self-report that assesses the 20 DSM-5 symptoms of PTSD. The PCL-5 was designed to be used by a variety of testers like healthcare and social sciences practitioners, non-professional researchers, students, and community members to determine the presence and severity of PTSD following an initial traumatic experience. It is intended for use in the clinical setting as a screening device for initial assessment of PTSD.

The PCL-5 test comprises 20 questions and the patient answer them by self-assessment of his PTSD symptoms over the past month, using a 5-point scale ranging from 0 to 4 with 0 meaning not at all and 4 when the symptom is extremely present. An overall symptom severity score with a range from 0 to 80 can be computed as a sum of the scores for each of the 20 items. According to preliminary research, samples with a PCL-5 cutoff score greater than 38 are likely to have PTSD. The test takes approximately 5–10 min to complete, and the interpretation is made by a clinician. The PCL has been translated into over 50 languages and is universally used in research studies, and clinical trials all over the world.

With the increasing number of patients with this mental condition, as well as the high cost of therapy sessions, has led to an increased interest in an alternative, more reliable tool that is accessible to anyone at anytime and that is easy to use for PTSD treatment and monitoring. Thus, we propose in this paper PTSDetection application for the diagnosis, the assistance and the follow-up of PTSD patients.

The paper is organised as follows. Related works to mental health disorders and telemedicine are discussed in Sect. 2. In Sect. 3, our new proposed methods for monitoring PTSD are presented. The different technologies used to develop the application and the different architectures are presented in Sect. 4. In Sect. 5 is consecrated for the application evaluation. Finally, Sect. 6 give the conclusion of this research work and discusses future works.

## **2 Related Work**

### ***2.1 Mobile and Web Applications for Health Monitoring and Surveillance***

The use of software applications by clinicians and healthcare professionals has changed many aspects of medical practice [21]. With the widespread use of devices in healthcare settings, the development of medical software applications has grown rapidly [1]. Several applications are now available to assist medical staff with many major tasks, such information and time management, recording and saving, communication and consultation, referral and information gathering, patient assistance and follow-up, help decision-making, and teaching and training [5]. Patients who require 24/7 monitoring have traditionally been monitored with fixed, bulky equipment that may restrict their mobility or not effectively provide monitoring over predetermined

periods of time. Increased mobility naturally results in a change in the user's contextual information, such as location and available resources around them. The use of mobile web services provides a natural opportunity for mobility and allows caregivers in remote locations to "keep an eye" on patients. Providing such remote health monitoring through mobile and web applications offers a low-cost alternative to traditional telemedicine approaches. The requirements are a device, a mobile web service hosted on the patient's device and a simple user interface on both sides and a standard web browser.

## ***2.2 Mental Disorders and Mental Health Telemedicine***

Telemedicine can provide help for mental health disorders and issues that are difficult to treat in person, such as addiction. With telemedicine, a healthcare provider can examine a patient remotely and diagnose and prescribe treatment. Telemonitoring is also used to help monitor the health of individuals who have lost the ability to communicate verbally or through other means.

In a recent study [19], medication adherence with telemedicine has been investigated for adult persons with severe mental disorders. In their clinical trial, patients with schizophrenia or bipolar illness were randomly chosen to either the intervention group or the usual care control group [19]. Trained nurses conducted the intervention. The Medication Adherence Report Scale was used to assess medication adherence. The study found that patients who received telemedicine intervention had improved medication adherence as compared to the control group. The telemedicine intervention was successful in improving medication adherence. The study found that patients who received telemedicine intervention had improved medication adherence as compared to the control group. In another work, a system for tracking people with bipolar disorder in the mental health field [9] was introduced. This system collects speech while performing tasks and includes a questionnaire system, sleep activity monitoring, and medicine intake monitoring capabilities. It is built on textile-based autonomic nervous sensor technology.

When stress becomes chronic, serious health issues result. The biosensors in mobile phones can detect psychological stress, but they cannot detect mental stress, which is determined by monitoring the heart rate and its variability. The R peaks and ECG data were processed using an algorithm, which provided good results [4]. Numerous issues affected the management of chronic illnesses in both home and hospital settings. Different parameters were chosen to provide the optimum care for the patients, despite the fact that the monitoring and abnormality detection methods differed between hospital and home. To integrate the two settings, a new architecture was created [8]. The heart rate variability is used to measure human emotions and stress. The ECG alterations that corresponded to the four different moods were recorded. The optimal emotional states were determined using the physiological characteristics [22]. Under tough circumstances, the biomedical signals were used to monitor the physiological parameters. The condition was discovered using micro

sensors and algorithms [18]. The mood of the person changes since stress has been rising among people at all levels. A mood recognition system based on the frequency of use of several mobile applications was also introduced [10].

Clinical depression or also called Major Depressive Disorder (MDD) is the most studied mental disorder by the computer vision community and several machine learning-based solutions have been proposed based on the video data [11, 11–17]. The proposed computer-aided diagnostic systems for continuous assistance of patients with Major Depressive Disorder are based on high performing deep neural networks-based approaches and they have demonstrated that video is a fast, non-invasive and non-intruded approach for depression recognition and surveillance and it is convenient for real-world applications. More recently, in [15, 16], a proposed clinical decision support systems that uses audiovisual cues extracted from the video recordings of clinical interviews in order to predict depression relapse and guide treatment decisions. The proposed clinical decision support systems uses machine learning and behavioral profiling techniques to extract audiovisual cues from the video recordings of clinical interviews and to predict depression relapse [15, 16].

### ***2.3 Mobile Application for Post-traumatic Stress Disorder Diagnosis and Follow-Up***

Despite the adequacy of current treatments for PTSD, access to treatment, particularly for veterans, continues to pose critical challenges that can impact adherence to appropriate treatment [7]. First, access to mental health care is becoming increasingly needed as people have become more conscious of their stress and anxiety levels and as the health sector suffers from a shortage of medical staff. Mental health offices continue to be understaffed and despite the growing demand for mental health care. Many people with post-traumatic stress disorder (PTSD) are unable to get to health care jurisdictions for treatment due to social, cultural and access facility barriers.

New health care interventions that rely on electronic information and communication processes, such as the use of mobile applications, could be a way to overcome several barriers for people who are unable or unwilling to access mental health care. Over the past decade, computer-based applications (such as cell phone alcohol app mediations and machine learning approaches to distinguish PTSD) have been used to extend reach, provide real-time verification, and offer a more comprehensive treatment pathway through advanced support steps. A subsequent study of veterans treated for PTSD showed that those with access to a mobile device were eager to use a versatile wellness intervention through apps to monitor their health condition [6].

### **3 The Mobile Application**

#### ***3.1 Video and Data Acquisition***

Data acquisition is the process of collecting data. In this project, the application collects several data for further analysis and statistics. The acquired data represents demographic data that provides a better understanding of certain background characteristics of an audience, whether it is age, ethnicity, sex and job. The user's answers will be given and recorded while the camera is open. The recorded video of him answering the questions will be saved and can be used later for statistical purposes.

#### ***3.2 Virtual Interviewer***

A virtual interviewer is designed in our proposed mobile application to play the role of the clinician. The virtual interviewer asks the user the 20 questions of the PCL-5 test to determine if he is suffering from post-traumatic stress disorder or not and to determine the PCL-5 score for this test. This score will be saved in the patient's history with the corresponding date. The design of a virtual human interviewer aims to engage the patient in a face-to-face virtual interaction where he can feel more comfortable than in clinical interview. Besides, it creates interaction conditions for automatic assessment of distress indicators.

The virtual interviewer has several advantages: it promotes social distancing due to Covid19 [20], it can solve the problem of the lack of medical and paramedical staff [2], and it helps the patient be more comfortable when answering the mental health questionnaires [3]. In fact, it has been noticed that people are more comfortable talking to a virtual agent or responding anonymously. One study indicated that following a combat deployment, the sub-sample of benefit individuals who namelessly replied the schedule PDHA side effect checklist detailed twofold to fourfold higher mental wellbeing side effects and the next intrigued in accepting care compared to the in general comes about inferred from the standard organization of the PDHA, which is identifiable and connected to benefit members' military records [23].

#### ***3.3 PCL-5 Questionnaire***

The PCL-5 can be a self-screening device to assess PTSD. It is used to conclude the existence of PTSD; a conclusive determination can be given by a properly trained clinician. It is a 20-item self-report degree that evaluates the 20 DSM-5 symptoms of PTSD. It has a variety of purposes, including: tracking symptom progression during and after treatment, screening for PTSD, and making a provisional diagnosis of PTSD. This tool can be used multiple times after diagnosis to assess the progression

of PTSD symptoms over time. A 5-point decrease has been proposed to reflect a solid lessening in side effects, meaning that the alter is not arbitrary. This can be used to test whether a person's symptoms are responding to treatment. A reduction of 10–20 points reflects a clinically significant change. In our proposed PTSDetection application, the PCL-5 questionnaire is implemented.

### ***3.4 Patient History Recording***

The history of the user is a typical approach to infer information about the user. An initial history of information must be produced when the user is registered. Each time a user completes the test, the results are displayed in a table and graph, and the application uses the information to assess whether or not the user has PTSD. After taking the test and answering the 20 questions, the application will determine whether or not the user has PTSD. The result will be stored in the database as a patient history. The PCL-5 test result is used to start a clinical history and medical record for the user.

## **4 Application Development**

In this section we will be discussing the general architecture of the application and the technological choices.

### ***4.1 General Architecture***

The physical architecture of our application is composed of 3 levels, called 3-tier architecture. First, we have the client layer which is mainly a web browser that represents the communication tool between users and our application. Indeed, the user sends HTTP<sup>1</sup> requests and receives the response in JSON<sup>2</sup> format. Secondly, we have the application server layer that takes care of the business processes of the application, it contains a Web layer that manages the client's requests and sends them to the other layers to perform the necessary processing. And last of all, we have the database server layer which is a database containing the information related to the users of the application.

This choice of this architecture is justified by its flexibility for applying new technologies. It also guarantees increased security and protect the privacy of the patients which is very important for mental health applications. With a three-tier

---

<sup>1</sup> Hypertext Transfer Protocol.

<sup>2</sup> JavaScript Object Notation.

architecture, access to the database is provided only by the application server. This server is the only one that knows how to connect to the database. It does not share any information that allows access to the data, including the database login and password. Furthermore, it provides better performance, given the sharing of tasks between different servers. Also, it significantly reduces deployment and administration costs. In fact, the main advantage of a three-tier architecture is the ease of deployment. The application itself is deployed only on the server part (application server and database server). The client itself can access it using the browser that is compatible with our application and is installed by default on all machines.

The presentation layer is developed with React Js and contains all the Man/Machine interfaces and services that ensure communication with the back-end. Indeed, React Js applications are modular applications. Each module is organized according to the following Hooks architecture: The component is a function that implements the business logic of the application and interacts with the views also called templates that describe HTML<sup>3</sup> pages. These interfaces provide services that are unique to the functionality of the application from which we use the back-end REST controller-level Web services. There is the security layer that ensures the security of the application thanks to the Web token JWT.<sup>4</sup> Also, there is the core layer. It contains all the core processes of our application. We then implement the different data processing. And last of all, the data persistence layer that contains the data handling logic.

## 4.2 *Technological Choice*

In the process of creating the project that would contain all of our features and after a thorough analysis of all the technologies, we chose ReactJs, Django and MongoDB-ATLAS. And in this section we have detailed the languages used by classifying them, according to their specificities, in three layers.

**Frontend** For the frontend we decided to use React.js. Its is an open-source library developed in JavaScript. It is widely used for creating user interfaces for single-page apps. It can handle the view layer for mobile/web applications. Another advantage of React, it allows to build reusable UI components.<sup>5</sup>

**Backend** The proposed application will integrate in the future deep neural networks for the diagnosis and the follow-up of PTSD. For interoperability reasons, django and django REST framework are chosen for the backend of the app. Django is a high-level Python web framework for fast development. In fact, Django REST framework: is a strong and flexible open source toolkit for building Web APIs.<sup>6</sup>

---

<sup>3</sup> HyperText Markup Language.

<sup>4</sup> JSON Web Tokens.

<sup>5</sup> <https://www.c-sharpcorner.com/article/what-and-why-reactjs/>.

<sup>6</sup> <https://www.django-rest-framework.org/>.

**Database** To ensure many users usage, MongoDB ATLAS and Djongo are chosen. MongoDB Atlas, a cloud management database system. It has several advantages of dealing with complexity of deploying/managing/healing the placing on the cloud service provider. It is in fact the best way to place and run MongoDB in the cloud.<sup>7</sup> Also, to make data manipulation more easy and to simplify constructing queries, we used Djongo. It is an integrated approach for database interfacing.<sup>8</sup>

**Development tools** The chosen editor is Visual Studio Code (VS Code). It integrates several features that facilitate development such as syntax highlighting, IntelliSense auto-completion system, source code management with simple and powerful Git integration, integrated terminal support that allows to select and use the Shell of the development platform.

And to test the functionalities of our APIs in the backend, we used Postman,<sup>9</sup> an API platform for creating and using APIs.

**Version control software** Git and Github are used for collaborative work environment. In fact, Git is a version control system that track change history and help to coordinate work among several work collaborators. And last of all, to store all the different versions of our project, we used Github.

## 5 Application Evaluation and Dataset Collection

The first step for the user is to register and log in. Once logged in, the user will have to turn on his camera to take the PCL-5 test. The subject will listen to the virtual interviewer and respond to each question vocally and select an answer by a click on the corresponding button. The application will determine if the user has PTSD based on their answers to 20 questions. Each question will have a response on a scale of 0–4. If the total scale of all the responses is greater than 38, then the subject is diagnosed with PTSD. The test procedure is illustrated step by step in Fig. 1.

For user satisfaction evaluation and for further future study of PTSD, PTSDetection app is used to collect a new dataset from 100 participants. All participants are normal control subjects and they did not present PTSD antecedents. Each test takes around 10 min and each participant is asked to rank the app and to give comments and recommendations for improving the app. The average score given by the participants is **4.2/5**. This score is satisfactory for app deployment for more collect of data from PTSD subjects in collaboration with Henri Mondor Hospital in Paris. The in-wild dataset collected from 100 subjects will be used for self-supervised learning of PTSD patterns from audiovisual cues.

---

<sup>7</sup> <https://www.mongodb.com/basics/mongodb-atlas-tutorial>.

<sup>8</sup> <https://www.djongomapper.com/>.

<sup>9</sup> <https://www.postman.com/>.



(a) Sign In



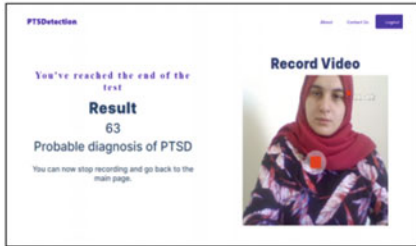
(b) Profile



(c) Ready For Test



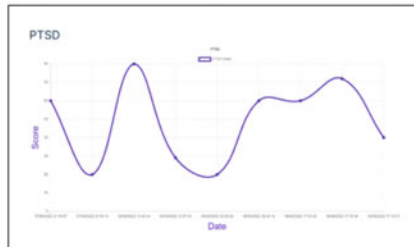
(d) Starting Test



(e) Test Result

PTSD Score	Date	PTSD Result
65	17/08/2022 21:50:07	Probable Diagnosis of PTSD
20	17/08/2022 21:51:13	Unable to take PTSD
65	28/08/2022 16:49:49	Probable Diagnosis of PTSD
29	28/08/2022 16:51:24	Unable to take PTSD
20	28/08/2022 16:59:20	Unable to take PTSD
65	28/08/2022 16:59:19	Probable Diagnosis of PTSD
65	28/08/2022 17:51:43	Probable Diagnosis of PTSD
75	28/08/2022 17:51:16	Probable Diagnosis of PTSD
45	28/08/2022 17:51:51	Probable Diagnosis of PTSD

(f) Table of Previous Tests and patient history



(g) Patient history based on the scores of the PCL-5 test.

Fig. 1 A screening of the proposed PTSDetect monitoring application



## 6 Conclusions and Future Work

In this paper, a new PTSD monitoring application is introduced. The proposed mobile application is based on the acquisition and the analysis of the responses of the PCL-5 test and the video data from the patients. By using this app, it is possible to monitor and to follow-up PTSD symptoms more accurately and at home. The proposed app can be used to recognize changes in patients' audiovisual cues and self reported symptoms and then it can alert the doctor and caregivers to potential risks. In future work, we are planning to propose a deep learning-based approach for PTSD recognition and to integrate it into the PTSDetect application. The automatic analysis of the video data of PTSD patients will give more information about patients' mental state. The aim is to get a better understanding of patients' mental health and be able to predict PTSD symptoms in them.



## References

1. Aungst, T.D.: Medical applications for pharmacists using mobile devices. *Annals of Pharmacotherapy* **47**(7–8), 1088–1095 (2013)
2. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (2010)
3. Baltrušaitis, T., Robinson, P., Morency, L.P.: 3d constrained local model for rigid and non-rigid facial tracking. In: *2012 IEEE conference on computer vision and pattern recognition*, pp. 2610–2617. IEEE (2012)
4. Carbonaro, N., Anania, G., Dalle Mura, G., Tesconi, M., Tognetti, A., Zupone, G., De Rossi, D.: Wearable biomonitoring system for stress management: A preliminary study on robust ECG signal processing. In: *2011 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, pp. 1–6. IEEE (2011)
5. Divall, P., Camosso-Stefinovic, J., Baker, R.: The use of personal digital assistants in clinical decision making by health care professionals: a systematic review. *Health informatics journal* **19**(1), 16–28 (2013)
6. Erbes, C.R., Stinson, R., Kuhn, E., Polusny, M., Urban, J., Hoffman, J., Ruzek, J.I., Stepnowsky, C., Thorp, S.R.: Access, utilization, and interest in mhealth applications among veterans receiving outpatient care for PTSD. *Military Medicine* **179**(11), 1218–1222 (2014)
7. Fortney, J.C., Pyne, J.M., Kimbrell, T.A., Hudson, T.J., Robinson, D.E., Schneider, R., Moore, W.M., Custer, P.J., Grubbs, K.M., Schnurr, P.P.: Telemedicine-based collaborative care for posttraumatic stress disorder: a randomized clinical trial. *JAMA psychiatry* **72**(1), 58–67 (2015)
8. Jeong, S., Youn, C.H., Shim, E.B., Kim, M., Cho, Y.M., Peng, L.: An integrated healthcare system for personalized chronic disease care in home-hospital environments. *IEEE Transactions on Information Technology in Biomedicine* **16**(4), 572–585 (2012)
9. Lanata, A., Valenza, G., Nardelli, M., Gentili, C., Scilingo, E.P.: Complexity index from a personalized wearable monitoring system for assessing remission in mental health. *IEEE Journal of Biomedical and Health Informatics* **19**(1), 132–139 (2014)
10. Ma, Y., Xu, B., Bai, Y., Sun, G., Zhu, R.: Daily mood assessment based on mobile phone sensing. In: *2012 ninth international conference on wearable and implantable body sensor networks*, pp. 142–147. IEEE (2012)
11. Muzammel, M., Othmani, A., Mukherjee, H., Salam, H.: Identification of signs of depression relapse using audio-visual cues: A preliminary study. In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 62–67. IEEE (2021)

12. Muzammel, M., Salam, H., Hoffmann, Y., Chetouani, M., Othmani, A.: Audvowelconsnet: A phoneme-level based deep CNN architecture for clinical depression diagnosis. *Machine Learning with Applications* **2**, 100,005 (2020)
13. Muzammel, M., Salam, H., Othmani, A.: End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. *Computer Methods and Programs in Biomedicine* **211**, 106,433 (2021). <https://doi.org/10.1016/j.cmpb.2021.106433>. <https://www.sciencedirect.com/science/article/pii/S0169260721005071>
14. Othmani, A., Kadoch, D., Bentounes, K., Rejaibi, E., Alfred, R., Hadid, A.: Towards robust deep neural networks for affect and depression recognition from speech. In: *International Conference on Pattern Recognition*, pp. 5–19. Springer (2021)
15. Othmani, A., Zeghina, A.O.: A multimodal computer-aided diagnostic system for depression relapse prediction using audiovisual cues: A proof of concept. *Healthcare Analytics* p. 100090 (2022). <https://doi.org/10.1016/j.health.2022.100090>. <https://www.sciencedirect.com/science/article/pii/S2772442522000387>
16. Othmani, A., Zeghina, A.O., Muzammel, M.: A model of normality inspired deep learning framework for depression relapse prediction using audiovisual data. *Computer Methods and Programs in Biomedicine* **226**, 107,132 (2022)
17. Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., Othmani, A.: Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control* **71**, 103,107 (2022). <https://doi.org/10.1016/j.bspc.2021.103107>. <https://www.sciencedirect.com/science/article/pii/S1746809421007047>
18. Rózanowski, K., Sondej, T., Lewandowski, J., Łuszczczyk, M., Szczepaniak, Z.: Multisensor system for monitoring human psychophysiological state in extreme conditions with the use of microwave sensor. In: *Proceedings of the 19th International Conference Mixed Design of Integrated Circuits and Systems-MIXDES 2012*, pp. 417–424. IEEE (2012)
19. Schulze, L.N., Stentzel, U., Leipert, J., Schulte, J., Langosch, J., Freyberger, H.J., Hoffmann, W., Grabe, H.J., van den Berg, N.: Improving medication adherence with telemedicine for adults with severe mental illness. *Psychiatric services* **70**(3), 225–228 (2019)
20. Speicher, M.: What is usability? a characterization based on ISO 9241-11 and ISO/IEC 25010. *arXiv preprint arXiv:1502.06792* (2015)
21. Wallace, S., Clark, M., White, J.: ‘it’s on my iphone’: attitudes to the use of mobile computing devices in medical education, a mixed-methods study. *BMJ open* **2**(4), e001,099 (2012)
22. Wang, C., Wang, F.: An emotional analysis method based on heart rate variability. In: *Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics*, pp. 104–107. IEEE (2012)
23. Warner, C.H., Appenzeller, G.N., Grieger, T., Belenkiy, S., Breitbach, J., Parker, J., Warner, C.M., Hoge, C.: Importance of anonymity to encourage honest reporting in mental health screening after combat deployment. *Archives of general psychiatry* **68**(10), 1065–1071 (2011)

# COVID-19 Diagnosis and Classification from CXR Images Using Vision Transformer



Md Mahbubur Rahman, Shihabur Rahman Samrat, Abdullah Al Ahad, Mahmud Elahi Akhter , Ibraheem Muhammad Moosa , Rajesh Palit , and Ashfia Binte Habib 

**Abstract** The COVID-19 pandemic is yet to come to a halt, and the current primary method of diagnosis is Reverse Transcription Polymerase Chain Reaction (RT-PCR). Although RT-PCR is reliable, it is known to have a long turnaround time and high false-negative rates that can severely hinder the accuracy of diagnosis. Alongside RT-PCR, Rapid Antigen Tests (RAT) are also used, but they have much lower accuracy than RT-PCR. Motivated by the flaws of the current diagnosis methods, we present a Vision Transformer-based classifier for the successful diagnosis and classification of COVID-19 using chest X-Ray (CXR) images. In order to address dataset imbalance and bias issues, a 15,000 sample CXR dataset was compiled, which consisted of 5000 CXR per class. Afterwards, a Vision Transfer (ViT) was fine-tuned on the dataset. Resnet-50 and DenseNet121 were used as baseline models. It is observed that for multiclass classification, the Vision Transformer-based model has the highest classification accuracy of 96.2% with a F1 score of 0.965 and the average precision and recall of 0.9617 and 0.962, respectively. This study demonstrates the adequacy of the ViT for the identification and classification of COVID-19 and Pneumonia.

**Keywords** COVID-19 · Vision transformer · Image classification

## 1 Introduction

The novel coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome (SARS), has emerged as the deadliest virus of the century resulting in about 367.8 million people infected with over 5.65 million deaths worldwide as

---

M. M. Rahman · S. R. Samrat · A. Al Ahad · R. Palit · A. B. Habib (✉)  
North South University, Dhaka, Bangladesh  
e-mail: [ashfia.habib@northsouth.edu](mailto:ashfia.habib@northsouth.edu)

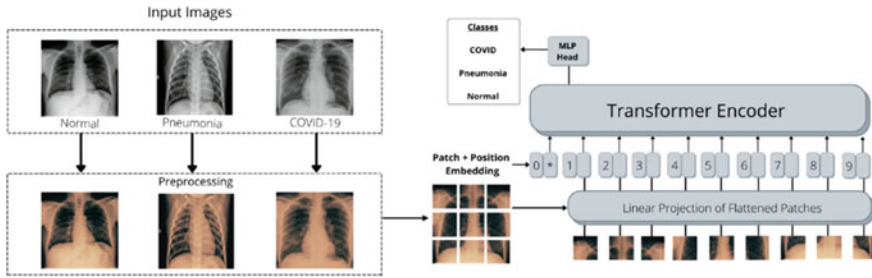
M. E. Akhter  
Università Degli Studi Dell'Aquila, 67100 L'quila, Italy

I. M. Moosa  
Temple University, Philadelphia, PA 19122, USA

of Jan. 28, 2022 [1]. The public health system is facing many obstacles in this pandemic. Shortage of medical resources increases the risk of infection for healthcare workers. Now, the primary method of diagnosis is Reverse Transcription Polymerase Chain Reaction (RT-PCR) or rapid antigen test. RT-PCR is considered the gold standard in diagnosing COVID-19 for its higher sensitivity and specificity than other primary diagnoses [2]. However, depending on the region and other factors, it can take several hours to days to get the result. Equipment to perform RT-PCR tests is also not adequate in all hospitals and health complexes. Before the COVID-19 pandemic, only a handful of tests required RT-PCR. Therefore, not all hospitals have the test available. It is a long process to equip these hospitals with these equipment's. Also, trained professionals are required to operate these equipment's. Creating newly trained professionals at an accelerated rate during this pandemic is not possible. For these reasons, the less economically developed countries are suffering to better detect COVID-19 cases. Positive radiological findings are present in the majority of the COVID-19 positive patients. Almost all hospitals and health complexes, even in less economically developed countries, have conventional radiographs or CXR machines. So, radiological diagnosis can be used for rapid screening of COVID-19, such as chest radiograph (CXR). Nevertheless, highly trained medical professionals are required to detect the disease, which is a scarce resource for many remote areas.

The use of deep learning approaches is widespread nowadays in different areas, and it also boosted the performance of many research fields. One essential application of deep learning in medical image analysis. CXR images of COVID-19 patients have distinct features that a deep learning-based model can classify. But at the same time, CXR images of COVID-19 patients and CXR images of Pneumonia patients can look exactly the same or have similar features. So, it is difficult to differentiate a COVID-19 CXR image from a Pneumonia CXR image. For these computer vision-related tasks, CNN has been widely used in the last few years. Almost all the image classification tasks were performed using CNN-based models in recent years for the high accuracy rate of CNN-based models. But like other technology, CNN also has some limitations due to its hard-inductive biases. Hard inductive biases restrict the capabilities of CNN when it handles a large amount of data [3]. In recent times, Vision Transformer has shown the capabilities of overtaking CNN in computer vision-related tasks. Recent studies show that Vision Transformer is comparable and, in some cases, better than the CNN-based model [4]. So, CXR or CT (CXR holds some practical advantages over CT) can be used for rapid screening by using vision transformers (ViT) techniques. But data is very important to complete these tasks. Decent amount of data from trustworthy sources are required for the building of Vision Transformer-based models that classify COVID-19 images effectively.

Roberts et al. [5] showed the various pitfall of deep learning-based classification of COVID-19. One of their primary concerns was the quality of COVID-19 datasets. Most of the public datasets were found to be unbalanced. They contained a small amount of COVID-19 data compared to other CXR data like normal pneumonia, lung cancer, etc. This kind of imbalance in data is as a bottleneck for classification performance of many standard learning algorithms. For instance, in [6] it was found that imbalanced dataset has inferior classification performance but using an improved



**Fig. 1** The proposed vision transformer model

SMOTE algorithm can give better result. A model trained with an unbalanced dataset can become very biased towards the class with the majority sample. If the data from different classes have huge differences in quantity, then the model may give biased output even if the accuracy is good. This can occur in misdiagnosis which can be disastrous and can affect the spread of COVID-19.

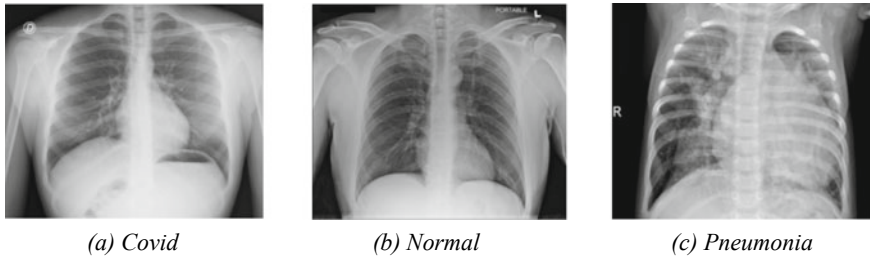
In order to address some of the concerns presented in [5], in this study we compile a custom dataset where the classes are balanced and other demographic are ensured without any form of implicit bias. Apart from that, this study also focuses on classification of COVID-19 from Pneumonia and healthy lungs and to evaluate the performance of Vision Transformer to classify CXR images. Therefore, the contribution of this paper is two folds,

1. We partially address the concerns of Roberts et al. [5] by compiling a balanced dataset without introducing any implicit bias.
2. Using the custom dataset, we analyze the performance difference between CNN and Vision Transformers for the task of COVID-19 detection.

The rest of the paper is as following, in Sect. 2, we discuss and illustrate related works. In Sect. 3, the proposed Vision Transformer model is discussed, and the entire pipeline is demonstrated. Section 4 gives an overview of the whole process of creating the datasets, data settings, and data preprocessing techniques. Section 5 summarizes the experiment process, analyzes the results, and compares the proposed model with CNN-based baseline models. Finally, we conclude with Sect. 6 (Fig. 1).

## 2 Related Work

In this section, we present works related to COVID-19 detection through deep learning. In [7], the study used vision transformer to detect COVID-19 from normal, COVID-19, and pneumonia patients CXR images. In the multiclass classification to detect covid from normal, covid, and pneumonia CXR images, they got an accuracy of 92% and an AUC score of 98%. Luz et al. [8] used a family of deep artificial neural networks based on the EfficientNet and achieved an overall accuracy of



**Fig. 2** Chest x-ray images

93.9%. Mondal et al. [9] also used vision transformer-based model to detect COVID-19 and the model was called ViTCOS-CXR, which achieved an accuracy of 96%. Parl et al. [10], used a vision transformer-based model on three different data sets of CXR images. These datasets were curated into three classes; normal, other infections (which includes bacterial pneumonia and tuberculosis), and COVID-19. They got an average accuracy of 86.4, 85.9, 85.2% and AUCs of 0.941, 0.909, 0.915 on three different datasets. Zhang et al. [11], used a two-step transfer learning pipeline and a deep residual network framework called COVID19XrayNet to detect COVID-19 from x-ray images. They integrated two novel layers into the popular ResNet32, i.e., feature smoothing layer (FSL) and feature extraction layer (FEL). They achieved an overall accuracy of 91.92% (Fig. 2).

### 3 Proposed Model

Transformer is widely used in natural language processing. But Dosovitskiy et al. [4] showed the potential of a Transformer based model for computer vision. They showed how an image can be compared with text and how to process an image with a transformer model to classify images. In this study, we utilize Vision Transformer to design a COVID-19 detection model. We used transfer learning and fine-tuned the model by training the model with our dataset. The pipeline for our model training is same as vision transformer. First, the input image goes through a preprocessing pipeline. After preprocessing, the images get converted into multiple flattened patches. The patches and their corresponding positional embedding are added to the transformer encoder as a sequence of data. Vision transformers recreate the visual structure from the training data and the vision transformer's self-attention layers enable it to integrate information globally to recreate the visual structure. Then the MLP head classifies the image into classes. Figure 1 shows the proposed Vision Transformer model.

## 4 Data

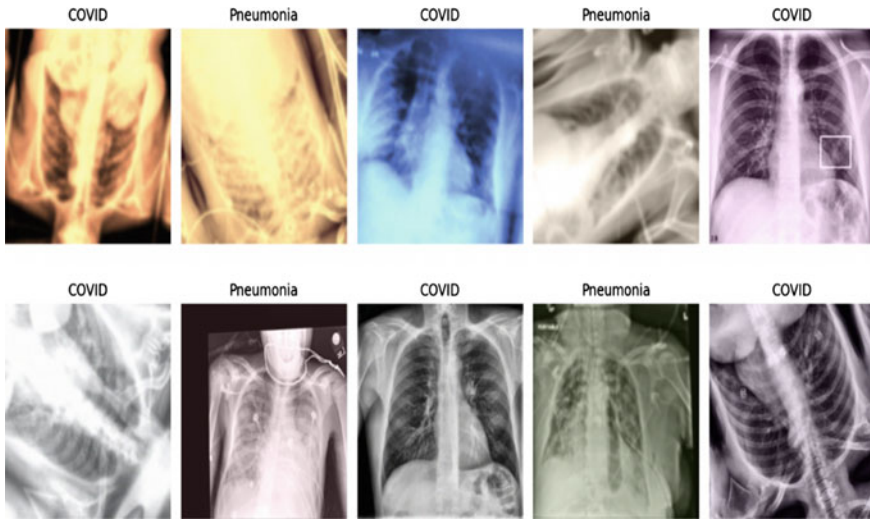
The dataset is one of the most important aspects in order to maximize a model’s performance. Finding image data, especially medical data like CXR (Chest X-Ray) is a challenging task by itself due to data privacy and rarity of some diseases. However, there are many public CXR data that are available online. But most of them have some form of issues with the data. There are high risks of biases in those data. Some data can be present in different datasets as these is public datasets. Some dataset only contains images with no other information, like how they verified the positive or negative data. In the dataset [12], they only included chest x-ray images of 1–5 years old babies. In the dataset [13], they included chest x-ray images of adult patients only. With these biased data, it is not possible to verify the effectiveness of any good model. Arranging private data from direct sources and verifying them with professionals is cost and time consuming. Therefore, the primary task was to find good sources of datasets and compiling from those sources to a single dataset. Before compiling the final dataset, we compiled several small datasets and did classification on them with no augmentation to verify that our compilation of the new datasets did not contain any implicit bias. We compiled a dataset of 15,000 CXR images with 5000 images from each class (COVID-19, Pneumonia, and Normal). We collected the data for our dataset from two public datasets that had very few limitations like other public datasets have, as mentioned above.

### 4.1 Dataset Details

The COVID-19 Radiography Dataset [14] contained 3615 COVID affected CXR images, 10,192 healthy CXR images, 6012 lung opacity affected CXR images, and 1345 pneumonia affected CXR images. On the other hand, the Extensive and Local Phase Enhanced COVID-19 X-Ray dataset [15] contained 8851 normal CXR images, 6045 CXR images of pneumonia affected patients, and 4038 CXR images of COVID affected patients. Out of these two datasets, we randomly sampled 5000 images for each class, as shown in the Table 1.

**Table 1** The arrangements of training, validation, and test datasets

Dataset/class	COVID	Normal	Pneumonia	Total
Train	3000	3000	3000	9000
Validation	1000	1000	1000	3000
Test	1000	1000	1000	3000
Total	5000	5000	5000	15,000



**Fig. 3** Chest x-ray images after preprocessing

## 4.2 Data Preprocessing

Various data augmentation techniques are applied to the training samples to improve the model's performance. For this purpose, we have used the OpenCV and Albumentations library. Since all the images in our dataset had a dimension of 299 by 299 pixels, we first resized the photos to a target size of 224 by 224 pixels using OpenCV. Furthermore, we applied normalization and CLAHE on our images. Then, using Albumentations, we applied ShiftScaleRotate with a shift and scale limit of 0.5 and rotate limit of 180. We also incorporated RGBShift with shift limits of 15. RandomBrightnessContrast and MultiplicativeNoise were also implemented. Next, we used HueSaturationValue with shift limits of 0.2. Finally, we added GaussianBlur and GaussianNoise (Fig. 3).

## 5 Experiments and Analysis

### 5.1 Implementation Details

For our experiments, we used Google Colab Notebooks for training and testing. For training, we used TPU instead of GPUs. We used Adam optimizer to train our model. The Adam-epsilon value for our model was  $1e-8$  to maintain numerical stability while updating the values. Weight decay was 0.01, and we used a learning rate of  $1e-4$ . The hyperparameter details are given in Table 2.



**Table 2** Training hyperparameters and data distributions

Parameters	Value
Adam epsilon	1e-8
Weight decay	0.01
Learning rate	1e-4
Training data	9000 images (60%)
Validation data	3000 images (20%)
Testing data	3000 images (20%)
Batch size (training, validation and testing)	30
Epochs	30
Steps	9000

### 5.2 Evaluation Method

Choosing an appropriate assessment metric is fundamental to overcome the bias among the differentiation of algorithms. For the classification standard, accuracy, review, F1 score, and exactness are the most common measures. Among all the identified samples, the number of correctly identified samples is called precision; from all positive samples, the number of correctly identified samples is called the recall. The harmonic average of precision and recall is called the F1 score. Accuracy is defined by the amount of correct classified samples from all samples.

- True-positive (TP):** The number of accurately labeled positive samples.
- True-negative (TN):** The correctly detected negative samples.
- False-positive (FP):** The number of negative examples classified as positive.
- False-negative (FN):** The number of positive cases classified as negative.

### 5.3 Results and Discussion

We trained our proposed vision transformer model ViT-M for 30 epochs. Then the system was evaluated over the test dataset of 3000 images, where we have 1000 images from each class. Our model could distinguish between the different classes with great accuracy. The test accuracy on our test dataset was 96.20%. The average precision of multiclass classification was 0.961, and the average recall of multiclass classification was 0.962. Our model performed better in classifying COVID-19 than other classes. As we can see in Fig. 5, the confusion matrix clearly shows us the performance of our model classifying COVID-19. Among 1000 COVID-19 chest x-ray images, our model successfully classified 988 images as COVID-19. There were only 13 false-negative cases. False-positive cases are even lower, only 8. As a result, our model achieved a precision of 0.988 and a recall of 0.992. Our model also achieved an AUC score of 99.15% in classifying COVID-19. Still, our model

can be seen getting confused sometimes in distinguishing between COVID-19 and other classes. In some mild cases of COVID-19, patients get less or no damage to the lungs. It can be a potential reason for this confusion. However, the model could accurately distinguish between COVID-19 and Pneumonia. We also trained a ViT called ViT-B with binary class that had only COVID-19 and healthy lungs. We can see that it achieves better performance compared to the multiclass model. However, the multiclass model is much more practical in many scenarios (Fig. 4; Table 3).

The accuracy and AUC score in classifying COVID-19 shows us that it performs much better than standard RT-PCR. It is also suitable for real-world deployment. It can be a great tool for assisting radiologists in reducing human errors. It can also be used as a primary diagnostic tool for COVID-19 detection for emergencies.

To evaluate the performance of our proposed vision transformer-based approach, we have compared its performance to two other classifiers: ResNet50 and DenseNet121. To achieve this, we trained both these models on the same dataset used in our vision transformer-based approach. For the ResNet50 model, we got an accuracy of 92.63%, which is lower compared to ViT-M. The confusion matrix in Fig. 5 shows that the ResNet-based model correctly classified 889 images out of 1000

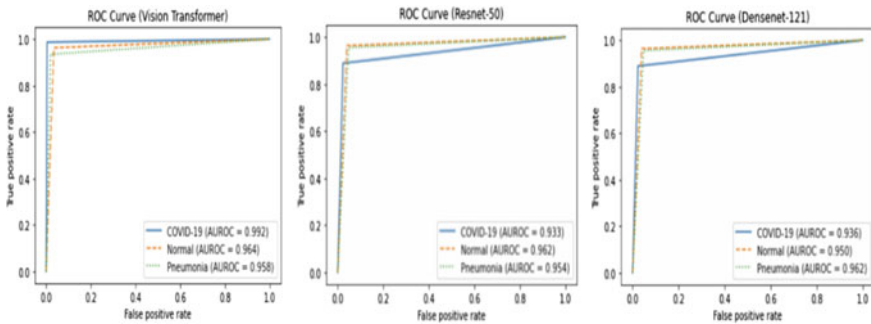


Fig. 4 ROC curve of each class for each model

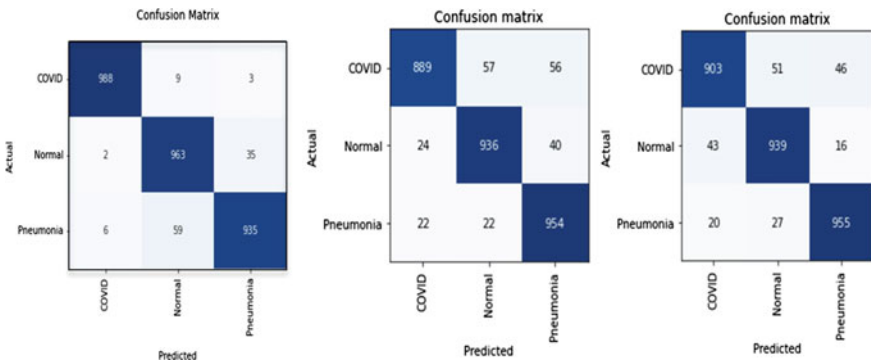


Fig. 5 Confusion matrix of vision transformer, ResNet-50, DenseNet-121

**Table 3** Evaluation of the proposed vision transformer model

Model	Accuracy	Precision	Recall	F1
ViT-M	96.20%	0.961	0.962	0.961
ViT-B	98.7%	0.988	0.992	0.990
Resnet-50	92.63%	0.916	0.926	0.921
Densenet-121	93.23	0.932	0.932	0.922

images for COVID-19, whereas ViT-M was able to successfully classify 988 images. In the case of the DenseNet121 model, on the same dataset, we got an accuracy of 93.23%, which is slightly better than that of ResNet50, but still falls behind our ViT-M model. In this case, as well, the DenseNet based model correctly classified 903 images (Fig. 5), which is still lower than our proposed approach. In both cases, ViT-M works better in comparison to the other two baseline models, especially at classifying COVID-19.

## 6 Conclusion

In this study, we proposed a vision transformer model to classify COVID-19 from CXR images. We also compiled a custom dataset to partially address the issues highlighted in [25]. We found that ViT outperformed ResNet50 and DenseNet121 in terms of COVID-19 detection. Upon comparison, our proposed model gave more promising results, especially in discriminating between COVID-19 and Pneumonia. This shows that for quick screening and diagnosis, ViT's can be a useful tool. It can also be used to assist doctors and radiologists as a second screening tool to reduce human errors.

## References

1. Worldometer, "Coronavirus toll update: Cases & deaths by country," Worldometers, 2021. <https://www.worldometers.info/coronavirus/>.
2. A. Tahamtan and A. Ardebili, "Real-time RT-PCR in COVID-19 detection: Issues Affecting the Results," *Expert Review of Molecular Diagnostics*, vol. 20, no. 5, Apr. 2020, doi: <https://doi.org/10.1080/14737159.2020.1757437>.
3. A. Goyal and Y. Bengio, "Inductive Biases for Deep Learning of Higher-Level Cognition," ResearchGate, Nov. 2020. [https://www.researchgate.net/publication/346555769\\_Inductive\\_Biases\\_for\\_Deep\\_Learning\\_of\\_Higher-Level\\_Cognition](https://www.researchgate.net/publication/346555769_Inductive_Biases_for_Deep_Learning_of_Higher-Level_Cognition).
4. A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv:2010.11929 [cs]*, Oct. 2020, [Online]. Available: <https://arxiv.org/abs/2010.11929>

5. Roberts, M., Driggs, D., Thorpe, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell* 3, 199–217 (2021). <https://doi.org/https://doi.org/10.1038/s42256-021-00307-0>
6. S. Wang, Y. Dai, J. Shen, and J. Xuan, “Research on expansion and classification of imbalanced data based on SMOTE algorithm,” *Scientific Reports*, vol. 11, no. 1, p. 24039, Dec. 2021, doi: <https://doi.org/10.1038/s41598-021-03430-5>.
7. D. Shome et al., “COVID-Transformer: Interpretable COVID-19 Detection Using Vision Transformer for Healthcare,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 21, p. 11086, Oct. 2021, doi: <https://doi.org/10.3390/ijerph182111086>.
8. E. Luz et al., “Towards an effective and efficient deep learning model for COVID-19 patterns detection in X-ray images,” *Research on Biomedical Engineering*, Apr. 2021, doi: <https://doi.org/10.1007/s42600-021-00151-6>.
9. A. K. Mondal, A. Bhattacharjee, P. Singla, and P. Ap, “xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography,” [www.techrxiv.org](http://www.techrxiv.org), Jul. 2021, doi: <https://doi.org/10.36227/techrxiv.14912367.v1>.
10. S. Park et al., “Vision Transformer for COVID-19 CXR Diagnosis using Chest X-ray Feature Corpus,” *arXiv:2103.07055 [cs, eess]*, Mar. 2021, [Online]. Available: <https://arxiv.org/abs/2103.07055>
11. R. Zhang et al., “COVID19XrayNet: A Two-Step Transfer Learning Model for the COVID-19 Detecting Problem Based on a Limited Number of Chest X-Ray Images,” *Interdisciplinary Sciences: Computational Life Sciences*, vol. 12, no. 4, pp. 555–565, Sep. 2020, doi: <https://doi.org/10.1007/s12539-020-00393-5>
12. P. Mooney, “Chest X-Ray Images (Pneumonia),” *kaggle.com*. <https://www.kaggle.com/paulti mothymooney/chest-xray-pneumonia>.
13. E. Bilello, “Medical Imaging Data Resource Center (MIDRC) - RSNA International COVID-19 Open Radiology Database (RICORD) Release 1c - Chest x-ray Covid+ (MIDRC-RICORD-1c) - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki,” [wiki.cancerimagingarchive.net](http://wiki.cancerimagingarchive.net), Mar. 05, 2021. <https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70230281>
14. T. Rahman, Dr. M. Chowdhury, and A. Khandakar, “COVID-19 Radiography Database,” *kaggle.com*. <https://www.kaggle.com/tawsifurrahman/covid19-radiography-database>.
15. endiqq, “Extensive and Local Phase Enhanced COVID-19 X-Ray,” *kaggle.com*. <https://www.kaggle.com/endiqq/largest-covid19-dataset>.
16. M. Tamal, M. Alshammari, M. Alabdullah, R. Hourani, H. A. Alola, and T. M. Hegazi, “An integrated framework with machine learning and radiomics for accurate and rapid early diagnosis of COVID-19 from Chest X-ray,” *Expert Systems with Applications*, vol. 180, p. 115152, Oct. 2021, doi: <https://doi.org/10.1016/j.eswa.2021.115152>.
17. T. N. Minh, M. Sinn, H. T. Lam, and M. Wistuba, “Automated Image Data Preprocessing with Deep Reinforcement Learning,” *arXiv:1806.05886 [cs]*, Apr. 2021, [Online]. Available: <https://arxiv.org/abs/1806.05886>.
18. A. Ulhaq, J. Born, A. Khan, D. P. S. Gomes, S. Chakraborty, and M. Paul, “COVID-19 Control by Computer Vision Approaches: A Survey,” *IEEE Access*, vol. 8, pp. 179437–179456, 2020, doi: <https://doi.org/10.1109/access.2020.3027685>.
19. S. H. Kassania, P. H. Kassanib, M. J. Wesolowskic, K. A. Schneidera, and R. Detersa, “Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning Based Approach,” *Biocybernetics and Biomedical Engineering*, vol. 41, no. 3, pp. 867–879, Jul. 2021, doi: <https://doi.org/10.1016/j.bbe.2021.05.013>.
20. P. R. A. S. Bassi and R. Attux, “A deep convolutional neural network for COVID-19 detection using chest X-rays,” *Research on Biomedical Engineering*, Apr. 2021, doi: <https://doi.org/10.1007/s42600-021-00132-9>.
21. K. Sivarama Krishnan and K. Sivarama Krishnan, “Vision Transformer based COVID-19 Detection using Chest X-rays,” *NASA ADS*, Oct. 01, 2021. <https://ui.adsabs.harvard.edu/abs/2021arXiv211004458S/abstract> (accessed Jan. 19, 2022).

22. M. N. A. Al-Hamadani and S. Suthaharan, "Evaluation of The Performance of Deep Learning Techniques," *ResearchGate*, Apr. 2015. [https://www.researchgate.net/publication/329029397\\_Evaluation\\_of\\_The\\_Performance\\_of\\_Deep\\_Learning\\_Techniques](https://www.researchgate.net/publication/329029397_Evaluation_of_The_Performance_of_Deep_Learning_Techniques).
23. Z. Wang *et al.*, "Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays," *Pattern Recognition*, vol. 110, p. 107613, Feb. 2021, doi: <https://doi.org/10.1016/j.patcog.2020.107613>.
24. L. Torrey and J. Shavlik, "Transfer Learning," *Handbook of Research on Machine Learning Applications and Trends*, pp. 242–264, doi: <https://doi.org/10.4018/978-1-60566-766-9.ch011>.

# Improved Techniques for the Conditional Generative Augmentation of Clinical Audio Data



Mane Margaryan, Matthias Seibold, Indu Joshi, Mazda Farshad, Philipp Fürnstahl, and Nassir Navab

**Abstract** Data augmentation is a valuable tool for the design of deep learning systems to overcome data limitations and stabilize the training process. Especially in the medical domain, where the collection of large-scale data sets is challenging and expensive due to limited access to patient data, relevant environments, as well as strict regulations, community-curated large-scale public datasets, pretrained models, and advanced data augmentation methods are the main factors for developing reliable systems to improve patient care. However, for the development of medical acoustic sensing systems, an emerging field of research, the community lacks large-scale publicly available data sets and pretrained models. To address the problem of limited data, we propose a conditional generative adversarial neural network-based augmentation method which is able to synthesize mel spectrograms from a learned data distribution of a source data set. In contrast to previously proposed fully convolutional models, the proposed model implements residual Squeeze and Excitation modules in the generator architecture. We show that our method outperforms all classical audio augmentation techniques and previously published generative methods in terms of generated sample quality and a performance improvement of 2.84% of Macro F1-Score for a classifier trained on the augmented data set, an enhancement of 1.14% in relation to previous work. By analyzing the correlation of intermediate feature spaces, we show that the residual Squeeze and Excitation modules help the model to reduce redundancy in the latent features. Therefore, the proposed model advances the state-of-the-art in the augmentation of clinical audio data and improves the data bottleneck for the design of clinical acoustic sensing systems.

---

Mane Margaryan and Matthias Seibold equally contributing first authors in alphabetical order.

---

M. Margaryan · M. Seibold · I. Joshi · N. Navab  
Computer Aided Medical Procedures, Technical University Munich, Munich, Germany

M. Seibold (✉) · P. Fürnstahl  
Research in Orthopedic Computer Science, Balgrist University Hospital, University of Zurich,  
Zurich, Switzerland  
e-mail: [matthias.seibold@balgrist.ch](mailto:matthias.seibold@balgrist.ch)

M. Farshad  
Department of Orthopedics, Balgrist University Hospital, University of Zurich, Zurich,  
Switzerland

**Keywords** Generative neural networks · Data augmentation · Audio signal processing · Acoustic sensing · Computer aided medicine

## 1 Introduction

Medical acoustic sensing systems utilize air- and structure-borne acoustic signals that can be captured in a medical environment, such as vibration signals from surgical tools captured with contact microphones [1] or sounds acquired with air-borne microphones directly from the area of operation [2], to provide guidance and support in medical interventions and diagnostics. Because acoustic signals can be captured non-invasively, radiation-free, and the systems are low-cost and easy-to-integrate, acoustic sensing has great potential for the design of multimodal sensing paradigms for the support of human surgeons, surgical diagnostics, robotic surgery, or to analyze surgical workflow. Hereby, acoustic sensing can be used to obtain measurements for applications where conventional medical computer aided support systems are limited, for example for the assessment of implant-bone press-fit which is impossible to obtain using imaging or navigation [1, 2], or to complement the limitations of medical imaging for the assessment of implant loosening [3] or cartilage degeneration [4].

Exemplary applications for the successful application of acoustic sensing in medical interventions are error prevention in orthopedic surgery by analyzing drill vibrations to detect drill breakthrough [5], the evaluation of implant seating during insertion of the femoral stem component in Total Hip Arthroplasty (THA) [1, 2], or the guidance of the insertion process of surgical needles using structure-borne acoustic signals acquired from the distal end of the medical device [6]. Also in medical diagnostics, acoustic signals have been successfully employed, e.g. for cough detection [7] or the examination of heart sounds [8].

In the recent years, deep learning-based analysis methods have outperformed classical signal processing and machine learning techniques for the processing of acoustic signals [9] which has also been applied in the medical domain in first use cases and showed promising performance improvements [1, 5]. While these methods are very powerful, they require large-scale high-quality training data to achieve superior performance and generalization to unseen cases. One of the main challenges for medical applications, however, is the limited availability of large amounts of data due to the limited access to the real surgical environment, expensive acquisition of realistic data, and clinical requirements and regulations. While in the non-medical domain of audio deep learning research, large-scale audio datasets, such as the Librispeech dataset for speech recordings [10] or the UrbanSound-8K dataset for environmental audio [11], are publicly available, the medical domain is lacking large-scale community data for the development of medical acoustic sensing systems. Therefore, especially in the medical domain, data augmentation is a valuable tool to artificially increase the size of a training data set to increase the diversity of training examples and stabilize the training process. To address this issue, we published a medical audio

dataset in a previous work which contains acoustic signals recorded in the real operating room during THA procedures which resemble typical surgical actions such as hammering, drilling, or sawing [12] and proposed a data augmentation method based on a conditional generative adversarial network.

However, we note that several studies report that deep networks tend to learn redundant features due to the huge model capacity [13–15]. Channel attention has been successfully exploited to model channel level dependencies and facilitate learning of less redundant features [16–18] and subsequently improved model performance. Motivated by these observations, in this paper, we demonstrate that due to the huge number of model parameters, conditional generative adversarial network (cWGAN-GP [12]) learns redundant features. To combat this, we introduce a channel-wise attention mechanism in the generator sub-network through the implementation of Squeeze and Excitation [16] block and residual skip connections [19]. We provide visualizations that signify the reduced redundancy and subsequently, improved quality of generated mel spectrograms samples quantified by a custom version of the Fréchet Inception Distance [20]. As a result, the present work advances the state-of-the-art in data augmentation for the emerging field of medical acoustic sensing and addresses the important issue of data limitations for medical deep learning-based systems.

## 2 Materials and Methods

### 2.1 Data Set, Preprocessing, and Benchmark Augmentations

We use a publicly available data set<sup>1</sup> [12] recorded during real Total Hip Arthroplasty surgeries that contains sounds of the typical surgical actions that are performed during the intervention and resemble the different phases of the procedure. The data set includes 568 recordings with a length of 1–31 s and the following distribution:  $n_{raw,Adjustment} = 68$ ,  $n_{raw,Coagulation} = 117$ ,  $n_{raw,Insertion} = 76$ ,  $n_{raw,Reaming} = 64$ ,  $n_{raw,Sawing} = 21$ , and  $n_{raw,Suction} = 222$ .

We compute mel spectrograms, a feature representation for audio signals that obtains state-of-the-art results for deep learning-based audio signal processing systems [9], using non-overlapping sliding windows which results in the following sample distribution for the entire data set:  $n_{spec,Adjustment} = 494$ ,  $n_{spec,Coagulation} = 608$ ,  $n_{spec,Insertion} = 967$ ,  $n_{spec,Reaming} = 469$ ,  $n_{spec,Sawing} = 160$ , and  $n_{spec,Suction} = 899$ . Mel spectrograms provide a compact representation, capture time- and frequency-domain aspects about a signal and can be computed from a raw waveform by first computing the Short-time Fourier Transform (STFT)  $X$  and then filtering the resulting spectra using a triangular filter bank spaced evenly on the mel scale [21] to compute the mel spectrogram  $X_{mel}$ . All spectrograms computed within the present work have

---

<sup>1</sup> The data set can be obtained from: <https://rocs.balgrist.ch/open-access/>.



dimensions  $64 \times 64$  and are normalized using the formula  $X_{norm} = (X_{mel} - \mu)/\sigma$  where  $\mu$  is the mean and  $\sigma$  is the standard deviation computed over the entire data set.

A number of data augmentation techniques for acoustic signals have been proposed in prior research, among them classical raw signal based methods like adding noise, time stretching, and pitch shifting, as well as spectrogram-based methods, e.g. SpecAugment [22]. Furthermore, we compare the results of the proposed data augmentation framework with the results reported in our previous work [12] in which a standard convolutional conditional generative adversarial network with Wasserstein Loss with Gradient Penalty regularization [23] was employed.

## 2.2 Proposed Data Augmentation Method

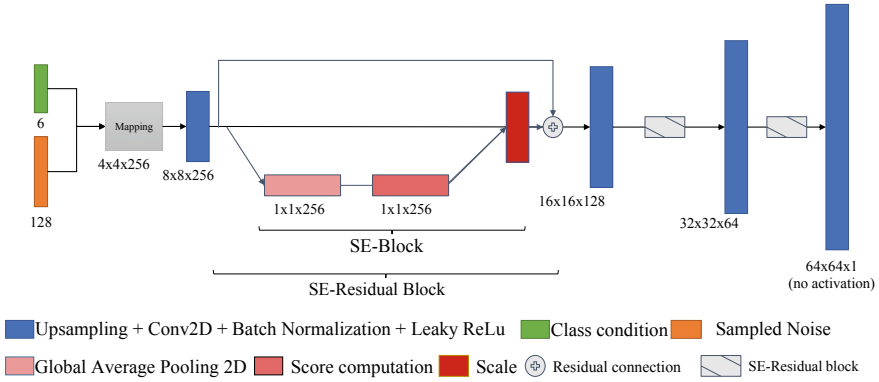
The architecture of the proposed GAN's generator is depicted in Fig. 1. It consists of 4 convolutional upsampling blocks followed by a squeeze-and-excitation block with a residual connection, a technique originally proposed in by Hu et al. [16]. The Squeeze and Excitation block consists of a global average pooling layer, which allows to *squeeze* global information to channel descriptors, a re-calibration part, which acts as a channel-wise attention mechanism and allows to capture channel-wise relationships in a non-mutually-exclusive way. The last operation scales the input's channels by multiplying them with the obtained coefficients. The Squeeze and Excitation mechanisms adds two fully connected layers with a ReLU activation function in between and a sigmoid function applied in the end as shown in Eq. 1.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1z)) \quad (1)$$

Here, the variables  $W_1$  and  $W_2$  have the dimensions  $(\frac{C}{r} \times C)$  and  $(C \times \frac{C}{r})$ , respectively,  $\sigma$  is the sigmoid function and  $\delta$  refers to a ReLU activation. The value of  $r$  is a hyperparameter and for our method it was chosen equal to 16 in an empirical manner.

The generator has an overall of 1,537,316 parameters. For the discriminator we use a fully convolutional network architecture with a total of 4,321,153 parameters analogous to our own previous work [12]. Both the generator and discriminator employ the LeakyReLU non-linear activation function throughout the whole network structure. As a loss function the Wasserstein Loss with Gradient Penalty (GP) was chosen with GP weight equal to  $\lambda = 10$ . For both the generator and the discriminator, we utilized the Adam optimizer with a learning rate of  $\lambda = 5 \times 10^{-4}$ . The discriminator was trained for 5 extra steps per epoch. The implementation and training of all reported results were done using Tensorflow/Keras 2.6 using a Google Cloud instance running a single NVIDIA T4 GPU.

The determination of when to stop the training process is notoriously difficult for the training of GANs. To assess the quality of the generated samples, we repeatedly compute a custom version of the Fréchet Inception Distance [20] which is computed



**Fig. 1** The schematic illustrates the structure of the proposed SE-ResNet generator for the generation of synthetic mel spectrograms

based on the features of the last convolutional layer of a ResNet-18 [19] pre-trained on the THA data set published in [12]. The training process is stopped when the lowest FID is observed which is computed using Eq. 2, where  $\mu_r$  and  $\mu_g$  is the feature-wise mean of the real and generated spectrograms,  $C_r$  and  $C_g$  are the covariance matrices.

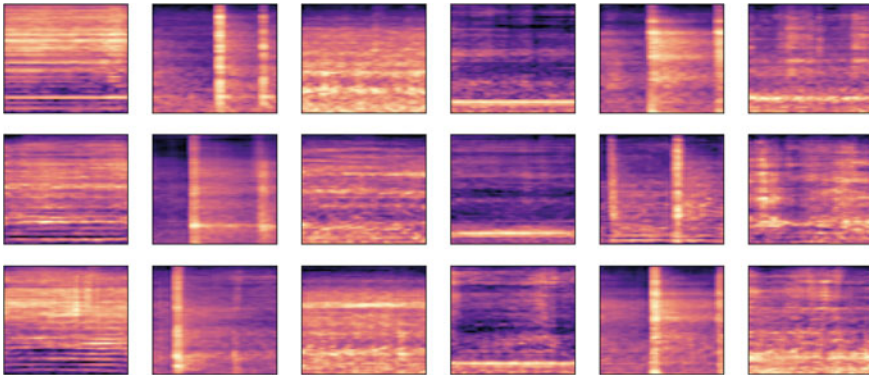
$$FID = \|\mu_r - \mu_g\|^2 + Tr(C_r + C_g - 2 * \sqrt{C_r * C_g}) \tag{2}$$

### 2.3 Classifier for Evaluation

For the evaluation of the proposed improved data augmentation method we employed a ResNet-18 classifier as previously reported in [12] which is a standard convolutional neural network architecture for spectrogram-based audio classification tasks and has been shown to achieve state-of-the-art results in medical acoustic sensing applications [1, 12]. To be able to compare the results presented within this work with the previous results, we augment 100% synthetic samples for each class present in the data set. The classifier was trained for 20 epochs using 5-fold cross-validation technique. We used categorical cross-entropy loss with the Adam optimizer and the following hyperparameters: learning rate =  $10^{-5}$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.99$ .

## 3 Results

In order to visually compare the quality of the proposed model, we present per-class randomly selected ground truth data, generated spectrograms from the pro-



**Fig. 2** Log-mel spectrogram of random samples for each class (top row); log-mel spectrogram of random generated images of our proposed model (second row); log-mel spectrogram of the model proposed in the previous work [12] (bottom row). Respective classes from left to right: Sawing, Adjustment, Reaming, Coagulation, Insertion, Suction

**Table 1** Comparison of different augmentation methods for clinical audio data

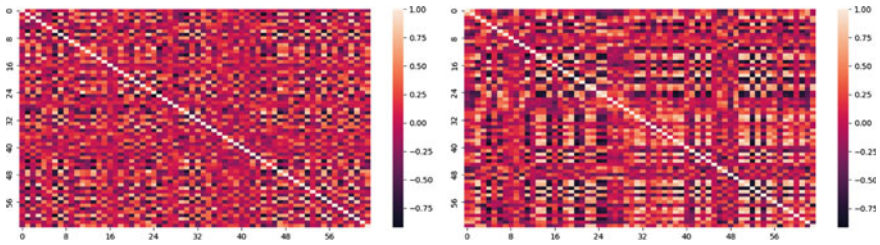
Augmentation method	FID	Macro F1-score (mean $\pm$ std, %)	Relative improvement (%)
No augmentation		93.9 $\pm$ 2.5	
White noise		92.87 $\pm$ 0.99	-1.03
Pitch Shift		94.73 $\pm$ 1.28	0.83
Time Stretch		95.0 $\pm$ 1.49	1.1
SpecAugment [22]		94.23 $\pm$ 1.14	0.33
cWGAN-GP [12]	3.30	95.60 $\pm$ 2.6	1.7
<b>Our method</b>	<b>3.01</b>	<b>96.74 <math>\pm</math> 1.03</b>	<b>2.84</b>

All reported results were obtained by applying the respective augmentation method to double the number of samples for each class of the public THA sounds data set

posed model, and synthetic spectrograms generated from the previous augmentation framework [12] in Fig. 2.

We stopped the training by frequently monitoring the quality of the generated samples through the computation of the FID as described in Eq. 2 and selected the best model with the lowest FID score which was subsequently used to augment the data set by doubling the number of samples for each class, the best performing augmentation strategy identified in previous work. We report the mean Macro F1 score over a five-fold cross validation experiment in the format mean  $\pm$  standard deviation. A comparison between the classifier performance with no augmentations, using classical signal- and spectrogram-processing-based methods, the method proposed in our own previous work [12], and the proposed model is shown in Table 1.

To analyze the redundancy in learned feature space of the proposed model and compare it with the previously published method, we plot the correlation matrices



**Fig. 3** Sample correlation matrices of features learned by the proposed model (left column) and cWGAN-GP published in previous work [12]. The correlation matrices are computed from an intermediate layer of the generator network. The plots represent the correlation in the feature space after the second-last convolutional layer with dimensions  $32 \times 32 \times 64$ . The significantly lower correlation values obtained after introducing Squeeze and Excitation block demonstrate the reduced correlation among features and therefore reduced feature redundancy

computed from intermediate layers of the network to analyze the redundancy of features in Fig. 3. The results show that the redundancy of features is significantly reduced by introducing residual Squeeze and Excitation modules in the generator network.

## 4 Discussion

Deep learning-based acoustic sensing has been shown to have high potential for clinical applications in diagnostics and interventional guidance, can be used for multimodal sensing to complement established assistance systems, and provide data beyond the limits of computer aided diagnostic and interventional support systems. However, to achieve state-of-the-art results, learning-based systems rely on big training data sets to generalize well for unseen cases. Obtaining these large amounts of clinical data is a common problem for the design of deep learning-based support and guidance systems in medicine. Advanced augmentation methods have been designed for medical imaging applications [24, 25] and a first method for the augmentation of clinical audio data sets has been proposed by the authors in previous work [12].

In the present work, the results show that the proposed method outperforms all previously suggested augmentation methods. In comparison to the first generative modeling based method for clinical audio data, we outperform the model by a margin of 1.14% in Macro F1-Score. While this is an incremental improvement, we could significantly improve the results by only adding a total number of 11,232 additional parameters which corresponds to a parameter growth of only 0.74% for the generator model. Furthermore, the correlation analysis of intermediate latent features revealed that the introduced residual Squeeze and Excitation modules reduce the redundancy in the learned features of the generator model. Therefore, the proposed architecture is a highly valuable extension in the generator architecture for an improved synthetic

generation of mel spectrograms. An improvement of 0.3 in the reported FID score underlines the capabilities of the proposed architectural modifications.

The proposed approach can generate any arbitrary number of samples for the classes present in the learned data set distribution and could therefore be employed to address data imbalance issues. However, in the current work we focused on improving the quality of the generated samples. Therefore, a more thorough investigation regarding the influence of different augmentation schemes using conditional generative data augmentation should be subject to future research.

By introducing a generative deep-learning method, the processing time for generating the augmentations increases in comparison to simple signal processing-based approaches. To investigate the capabilities of the proposed method, the model should be trained and evaluated on multiple relevant clinical audio data sets in future research.

## 5 Conclusion

In this work, we propose an enhanced generator architecture for conditional generative learning-based data augmentation of clinical audio data. We outperform all previously published methods and provide an in-depth analysis of the proposed modifications in the generator structure. The method is able to increase the quality of synthetically generated samples by 0.3 in terms of FID score and improves the performance of a classifier trained on the augmented data set by a margin of 2.84% in terms of Macro F1-Score. All presented results are evaluated on a public data set containing sounds of a Total Hip Arthroplasty procedure which was recorded in the real operating room and evaluated using a 5-fold cross validation scheme. The obtained results show that the proposed method has great potential to improve the problem of data limitations for the design of clinical acoustic sensing systems.

**Acknowledgements** This work is part of the SURGENT project under the umbrella of Hochschulmedizin Zürich.

## References

1. M. Seibold, A. Hoch, D. Suter, M. Farshad, P. O. Zingg, N. Navab, P. Fürnstahl, Acoustic-based spatio-temporal learning for press-fit evaluation of femoral stem implants, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2021, pp. 447–456.
2. Q. Goossens, L. Pastrav, J. Roosen, M. Mulier, W. Desmet, J. Vander Sloten, K. Denis, Acoustic analysis to monitor implant seating and early detect fractures in cementless THA: An in vivo study, *Journal of Orthopedic Research* (2020).
3. A. Arami, J.-R. Delaloye, H. Rouhani, B. M. Jolles, K. Aminian, Knee implant loosening detection: A vibration analysis investigation, *Annals of Biomedical Engineering* 46 (2018) 97–107.

4. K. S. Kim, J. H. Seo, J. U. Kang, C. G. Song, An enhanced algorithm for knee joint sound classification using feature extraction based on time-frequency analysis, *Computer Methods and Programs in Biomedicine* 94 (2) (2009) 198–206.
5. M. Seibold, S. Maurer, A. Hoch, P. Zingg, M. Farshad, N. Navab, P. Fürnstahl, Real-time acoustic sensing and artificial intelligence for error prevention in orthopedic surgery, *Scientific Reports* 11 (2021).
6. A. Illanes, A. Boese, I. Maldonado, A. Pashazadeh, A. Schaufler, N. Navab, M. Friebe, Novel clinical device tracking and tissue event characterization using proximally placed audio signal acquisition and processing, *Scientific Reports* 8 (2018).
7. K. S. Alqudaihi, N. Aslam, I. U. Khan, A. M. Almuhaideb, S. J. Alsunaidi, N. M. A. R. Ibrahim, F. A. Alhaidari, F. S. Shaikh, Y. M. Alsenbel, D. M. Alalharith, H. M. Alharthi, W. M. Alghamdi, M. S. Alshahrani, Cough sound detection and diagnosis using artificial intelligence techniques: Challenges and opportunities, *IEEE Access* 9 (2021) 102327–102344.
8. N. Giordano, M. Knaflitz, A novel method for measuring the timing of heart sound components through digital phonocardiography, *Sensors* 19 (8) (2019).
9. H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-y. Chang, T. Sainath, Deep learning for audio signal processing, *IEEE Journal on Selected Topics in Signal Processing* 14 (2019) 206–219.
10. V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An asr corpus based on public domain audio books, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206–5210.
11. J. Salamon, C. Jacoby, J. P. Bello, A dataset and taxonomy for urban sound research, in: 22nd ACM International Conference on Multimedia (ACM-MM'14), Orlando, FL, USA, 2014, pp. 1041–1044.
12. M. Seibold, A. Hoch, M. Farshad, N. Navab, P. Fürnstahl, Conditional generative data augmentation for clinical audio datasets, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2022, pp. 345–354.
13. J. Liu, B. Zhuang, Z. Zhuang, Y. Guo, J. Huang, J. Zhu, M. Tan, Discrimination-aware network pruning for deep model compression, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
14. I. Joshi, A. Utkarsh, P. Singh, A. Dantcheva, S. D. Roy, P. K. Kalra, On restoration of degraded fingerprints, *Multimedia Tools and Applications* (2022) 1–29.
15. P. Singh, V. K. Verma, P. Rai, V. Namboodiri, Leveraging filter correlations for deep model compression, in: Proceedings of the IEEE/CVF Winter Conference on applications of computer vision, 2020, pp. 835–844.
16. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
17. R. Roy, I. Joshi, A. Das, A. Dantcheva, 3d CNN architectures and attention mechanisms for deepfake detection, in: Handbook of Digital Face Manipulation and Detection, 2022, pp. 213–234.
18. M. Choi, H. Kim, B. Han, N. Xu, K. M. Lee, Channel attention is all you need for video frame interpolation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 10663–10671.
19. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
20. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local Nash equilibrium, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, p. 6629–6640.
21. S. S. Stevens, J. Volkman, E. B. Newman, A scale for the measurement of the psychological magnitude pitch, *The Journal of the Acoustical Society of America* 8 (1937).
22. D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, SpecAugment: A simple data augmentation method for automatic speech recognition, *Interspeech* 2019 (2019).
23. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of Wasserstein GANs, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 5769–5779.

24. M. Tirindelli, C. Eilers, W. Simson, M. Paschali, M. F. Azampour, N. Navab, Rethinking ultrasound augmentation: A physics-inspired approach, in: *Medical Image Computing and Computer Assisted Intervention*, 2021, pp. 690–700.
25. H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, M. Michalski, Medical image synthesis for data augmentation and anonymization using generative adversarial networks, in: A. Gooya, O. Goksel, I. Oguz, N. Burgos (Eds.), *Simulation and Synthesis in Medical Imaging*, 2018, pp. 1–11.

# Learning from Failure: A Methodology for the Retrieve Stage of a Cardiovascular Case-Based Reasoning System



Ana Duarte  and Orlando Belo 

**Abstract** Finding the most suitable cases is a critical process in implementing a Case-Based Reasoning system. A poor choice of the selected case has a negative impact on all the subsequent steps in the Case-Based Reasoning cycle. Over the last years, several studies have focused on the objective of defining retrieval mechanisms for improving the Similarity-Based Retrieval process. However, these works do not integrate metrics that take into account how often a case has led to successful and unsuccessful solutions. In some areas, such as cardiovascular health, a solution that works for some individuals may lead to poor outcomes for others. In such situations, it is important that the cases in the case database accurately reflect the success ratio of each solution. Therefore, the retrieval process should integrate this measure. In this paper, we propose a new method for retrieving cases based on similarity and success. Thus, in addition to similarity, we establish new metrics that allow the average success of the case solutions in a case database to be taken into account.

**Keywords** Retrieve · Similarity measure · Case-based reasoning · Well-being indexes · Cardiovascular health

## 1 Introduction

Human reasoning is an extremely complex cognitive process that is difficult to replicate. From memory and learning acquired through specific past examples in particular contexts, humans have the unique ability to infer new knowledge and new ways of solving problems [1–3]. As an example, we can imagine the scenario of a basketball game. Even if the court is different or the speed of the ball is constantly changing, players have the ability to generalize and quickly adapt to new circumstances.

---

A. Duarte · O. Belo (✉)  
ALGORITMI R&D Centre/LASI, University of Minho, Campus of Gualtar, 4710-057 Braga,  
Portugal  
e-mail: [obelo@di.uminho.pt](mailto:obelo@di.uminho.pt)

A. Duarte  
e-mail: [id9618@alunos.uminho.pt](mailto:id9618@alunos.uminho.pt)



Supported by the way humans learn and solve problems based on their previous experience, a new branch of *Artificial Intelligence* (AI) formally emerged in the early 1990s: *Case-Based Reasoning* (CBR). This paradigm enabled the development of more efficient systems that are able to mimic part of human reasoning. This type of reasoning is particularly relevant in areas where there is a strong subjective component, such as Health [4, 5].

One of the biggest public health problems in modern society is *Cardiovascular Disease* (CVD). Due to the high mortality rate, the causes of CVD have been a subject of scientific research over the years. The influence of lifestyle on cardiovascular health is undisputed. Nevertheless, although it is a widely researched area, knowledge about cardiovascular health is still far from being fully understood and it cannot be generalised to all people in the same way. In the area of nutrition, for example, there are countries that advocate a direct link between egg consumption and an increased cardiovascular risk, while others, such as Australia and the USA, have less restrictive guidelines [6].

CBR systems may be particularly suited to addressing the complexity and subjectivity inherent to cardiovascular health. If human reasoning can be transferred into machines, it is possible to create systems that can adapt to new problems and suggest effective actions to improve cardiovascular health based on previous successful experiences. Aamodt and Plaza made one of the most important contributions related to the CBR paradigm. In 1994, the authors presented the CBR cycle model, which consists of the processes *retrieval*, *reuse*, *revise* and *retain* [7]. Its circular shape allows CBR systems to continuously learn and improve their performance as more experiments are tested. This type of model can be especially useful in the context of cardiovascular health, as it enables to retrieve the past experiences (cases) that best represent a *New Problem* (NP) and reuse their solution. Therefore, retrieve is a critical step in the CBR cycle since an inadequate selection of the retrieved cases from the case base may lead to inappropriate solutions for the NP.

In order to select the best case that represents the NP, a methodology based on the similarity of the NP to each of the cases in the case base must be followed. Typically, this process is carried out using similarity metrics to determine the case from the case base that best reflects the target problem [4]. Since not all analysed attributes have the same relevance to the final solution, it is common to weight each attribute and assign a higher value to those that have a greater impact on the outcome. This strategy for finding the cases that are closest to the NP is called *Similarity-Based Retrieval* (SBR). When applied to cardiovascular health, this approach fails in one important aspect. Because it focuses only on finding the most similar cases with a NP, this strategy is blind to the success of the solutions associated with each case. In such situations, the most appropriated approach is to retrieve, among all the cases that have a high similarity to the NP, the one case whose solution has produced the best improvements in terms of cardiovascular health. Thus, our research provided the definition of a novel approach for the retrieval process of CBR systems aimed at improving cardiovascular health. To this end, in addition to the similarity measure, we propose the consideration of two new complementary metrics.

We organized the rest of this paper as follows. In Sect. 2 we give an overview of the retrieval process in some studies using CBR systems. In Sect. 3, we introduce the problem under analysis in more detail and describe the methodology used. Then, in Sect. 4, we demonstrate the importance of considering other measures besides similarity. Finally, we discuss the main conclusions and possible future work (Sect. 5).

## 2 Used Strategies to Retrieve Cases

Just as people use their memory to remember the solutions to a given problem, we can analogously use the case base to find the cases that have led to the best solutions in the past. Whenever a NP is analysed, it is very likely that the case base will not contain records that fully match the new case. Therefore, the main objective of the first phase of the CBR cycle is to find a case in the case base that is sufficiently similar to the problem to be solved. This way, its solution can be reused for the NP. However, this process is not always as linear as it appears and should not be viewed in such a simplistic way. Depending on the domain, the choice of the reference case must be carefully analysed, and the most similar case is not always the most appropriate one to serve as a reference for the NP. For this reason, it is important to reflect and find a set of criteria to select the most appropriate case from the case base.

The issue of retrieving cases has been addressed by different authors depending on the context of the problems to be solved. In addition to the similarity measure, some researchers recognise the importance of new measures or new strategies to improve the retrieving stage. One of these examples can be found in [8]. In 2001, the authors stated that in certain contexts it is important to combine similarity with diversity for retrieving cases. This is especially applicable in recommender systems or other applications that provide the user with more than one suggested solution to choose from. In turn, [9] uses the genetic algorithm in another study in 2008 to determine the optimal number of nearest neighbours to retrieve in situations where, instead of choosing just one case, several similar cases are selected. A year later, in [10] is emphasised the need to include gain and loss functions in the case retrieval process. In this study, gain and loss are directly related to the potential benefits and costs associated with each solution, respectively. In 2014, [11] proposed the use of association rules to complement the SBR approach. These rules make it possible to value the interaction between attributes and at the same time the similarity between the cases and the NP to be solved. Later, in 2016, [12] introduced a retrieval approach that uses class association rules to generate an optimum tree of frequent patterns. More recently, in 2020, [13] developed a retrieval strategy that takes into account the interaction between attributes, the relative importance of each attribute and the “attitudinal character” of decision makers. The approach established by these authors involves a redefinition of “similarity” considering both local and features similarities.

In general, these studies suggest interesting alternatives to the typical case retrieval approach based only on SBR. One aspect they have in common is that they all assume that the cases in the case base always represent successful solutions. However, in

many real-world problems, there is no guarantee that a solution that has worked well in the past will be equally effective in a new situation. When this happens, each case in the case base should reflect both the successes and failures associated with its solution. In these situations, we argue that an effective retrieval process should not only include the similarity with the cases in the case base. Rather, it should also consider additional measures to assess, for each case, the likelihood that its solution will lead to success and the corresponding expected value of success. For this purpose, in addition to the similarity measure, an evaluation index for the degree of success of the solutions (high, low or failure) must be considered. Furthermore, the percentage of successful situations that have arisen based on a particular solution should also be taken into account.

In this way, our work aims to propose a methodology for retrieving cases based on similarity and success. This approach applies to CBR processes that integrate solutions into the case base that are usually successful but that also sometimes lead to failures.

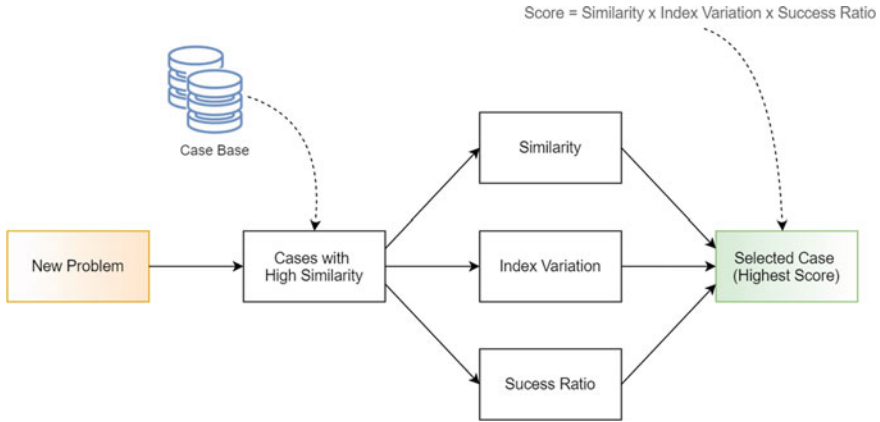
### 3 Other Measures Beyond Similarity

The case retrieval mechanism we propose in this paper applies to systems designed to promote cardiovascular health. Each of the cases in our case base consists of a set of characterising attributes, the proposed solution, and metrics to evaluate its success.

Two different approaches were considered for the selection of the description attributes of the case. On the one hand, we used some necessary attributes to quantify the value of the cardiovascular well-being index. On the other hand, we also selected a set of attributes that specifically characterise a person's lifestyle. In this work, the cardiovascular well-being index is a numerical measure expressed as a percentage that indicates the degree of cardiovascular health. To determine a person's well-being index, we adapted the QRISK2 calculation methodology [14]. QRISK2 is a tool recommended by several institutions for measuring cardiovascular risk and is widely used in the UK [15].

In brief, the method we developed for finding the most suitable case starts with applying the k-NN algorithm and selecting the  $k$  cases from the case base that are most similar to the NP. After calculating the similarity and determining the nearest neighbours, we set a minimum threshold for similarity to ensure that the cases that do not have a high similarity to the NP are excluded. In addition to the similarity measure, we also consider the values of the index variations associated with each case and the degree of success of their results, using Eq. 1. In the end, the case with the highest score becomes the reference for the NP (Fig. 1).

$$\text{Score} = \text{Similarity} \times \text{Index Variation} \times \text{Success Ratio} \quad (1)$$



**Fig. 1** Outline of the main steps of the proposed approach to case retrieval

Our methodology thus requires a deeper understanding of the concepts of similarity, index variation and success ratio. In the following subsections, we will address each of these points in detail.

### 3.1 Similarity

Similarity is a measure that can be calculated to retrieve cases in the case base whose attributes are similar to the NP. To determine it, we can use Eq. 2, considering the different local similarities and the weights of each attribute.

$$\text{Global similarity} = \frac{\sum_i w_i \times \text{Local similarity}_i}{\sum_i w_i} \tag{2}$$

The weighting of each attribute ( $w_i$ ) enables to associate the most important risk factors for cardiovascular health with the highest weighting values. At this level, we assumed that the attributes included in the formula for calculating the well-being index correspond to the most important risk factors. Therefore, these attributes were assigned the highest values of  $w_i$ . Depending on the type of attribute (binary, numeric or nominal), we specified different methods for determining the local similarities. For binary and numeric attributes, the Eqs. 3 and 4 in Table 1 were used.

In these equations,  $q$  represents the attribute value for the NP,  $x_i$  represents the attribute value for case  $i$  from the case base, and  $max$  and  $min$  represent the maximum and minimum allowable values for the analysed attribute.

For nominal attributes such as ethnicity, specific similarity tables were created based on the QRISK2 formula. For example, in terms of the risk of developing heart complications, QRISK2 indicates that the similarity between Caucasian and Indian

**Table 1** Methods used for calculating the local similarities depending on the attribute type

Attributes	Similarity
Binary attributes	$S = \begin{cases} 1, seq = x_i \\ 0, seq \neq x_i \end{cases} \quad (3)$
Numeric attributes	$S = 1 - \left( \frac{ q-x_i }{\max - \min} \right) \quad (4)$

**Table 2** Ethnicity scores for males and females according to the QRISK2 formula

Ethnicity	Men coefficient	Women coefficient
Caucasian	0.00	0.00
Indian	0.32	0.26
Pakistani	0.47	0.61
Bangladesh	0.52	0.34
Other Asian	0.14	0.15
Black African	-0.39	-0.18
Black Caribbean	-0.38	-0.35
Chinese	-0.41	-0.28
Other	-0.23	-0.16
Maximum range	0.93	0.96

ethnicities is greater than between Indian and Pakistani ethnicities. Therefore, to determine the similarity scores, we considered the QRISK2 coefficients associated with each ethnicity, as indicated in Table 2. To calculate this similarity, we used Eq. 4 to measure the closeness between the values of the coefficients used in the QRISK2 formula. As an example, the similarity value between a Caucasian man and an Indian man was calculated as follows:

$$\text{Similarity} = 1 - \left( \frac{|0.00 - 0.32|}{0.93} \right) = 1 - 0.34 = 0.66$$

A similar strategy was followed for the other nominal attributes.

### 3.2 Index Variation

The index variation works as a measure for evaluating the success of each solution. The cardiovascular well-being index is measured before and after following a solution and varies between 0 and 100. The index variation can thus be determined by Eq. 5.

$$\text{Index variation} = \frac{\text{Final index value} - \text{Initial index value}}{100} \quad (5)$$

Therefore, this means that higher values of the index variation correspond to more successful solutions. That is, a solution with an index variation of 10% means that, on average, each NP that followed that solution improved the cardiovascular well-being index by 10%.

### 3.3 Success Ratio

Another measure we propose to consider is the success ratio. This parameter corresponds to the percentage of successful outcomes relative to the total number of cases that followed a specific solution.

In the case base, we have included two fields related to the number of successes and the number of failures. Whenever a NP based on a case in the case base has a failure outcome, we increase the variable for the number of failures by one. Similarly, whenever there is a successful outcome, we increase the variable for the number of successes by one. Based on these considerations, we have defined Eq. 6, which allows us to determine the success level of each case in the case base.

$$\text{Success ratio} = \frac{\text{Number of successes}}{\text{Number of successes} + \text{Number of failures}} \tag{6}$$

The need to include this metric is related to the fact that there may be cases associated with a high index variation but that represent a significant number of failure outcomes. That is, there may be misleading situations where most of the problems solved with a given solution were not successful, but due to some outliers the value of the index variation is positive.

## 4 Evaluating Case Retrieval Strategies

In order to demonstrate the importance of considering the proposed measures, we can imagine the scenario shown in Table 3, where two cases with high similarity (above the defined threshold)—cases 1 and 2—are returned.

In this situation, the solution of case 1 increases the index value by 10%, on average. Case 2, on the other hand, leads to an average improvement of the index value by 9%. At first glance, i.e. without taking other parameters into account, case 1 seems to be the best reference, as its solution leads to a larger percentage increase

**Table 3** Example 1—comparison of the index variation of two cases

Case	Index variation
Case 1	0.10
Case 2	0.09

in the cardiovascular well-being index. Thus, applying its solution to the NP is more likely to also result in a larger increase in the value of the well-being index, which consequently translates into a more significant improvement in cardiovascular health. Index variation is thus a relevant factor to consider in the retrieval process. However, this is not the only important aspect to take into account.

Even if the returned cases already have a high similarity to the NP, there may be significant differences between their values. For example, one case may have a similarity of 95% to the NP and another case may have a similarity of 85%. In spite of the fact that both have high values, the first case is more similar to the NP and therefore it is more likely to lead to closer solutions and results. In this sense, similarity is a parameter that should not be neglected. Although it is already ensured that the returned cases have a high similarity to the NP, the most similar cases should be valued.

In addition to these factors, it is also necessary to quantify the ability of the cases to produce good results by means of a success ratio. As an example, we can consider two cases with the same similarity values with the NP. The first case has an average index variation of + 10%, while the second case has an average index variation of + 9%. In this scenario, without considering other parameters, the first of these cases would be the retrieved case due to its higher average index variation. Nevertheless, it is necessary to consider an important factor related to the probability of success of the case. This probability implies taking into account the number of records that led to success or failure. Let us assume that 30 records followed the solution of case 1 and 40 followed the solution of case 2, as shown in Table 4.

Although the index variation is, on average, higher in case 1, 6 of the 30 records that followed this case are unsuccessful cases. On the other hand, the records that followed the solutions of case 2 were always successful and increased the index value, although in a less visible manner comparing to case 1. Thus, according to Eq. 6, the first case has a success rate of 80% and the second a success rate of 100%. In situations like this, it is not clear which case should be chosen. However, by applying Eq. 1, it is possible to calculate a score that takes the three proposed measures into account (Table 5).

**Table 4** Example 2—comparison of the success ratio of two cases

Case	Number of failures	Number of successes	Total	Index variation	Success ratio
Case 1	6	24	30	0.10	0.80
Case 2	0	40	40	0.09	1.00

**Table 5** Example 3—combined comparison of the similarity, index variation and success ratio of two cases

Case	Similarity	Index variation	Success ratio	Score
Case 1	0.90	0.10	0.80	0.072
Case 2	0.90	0.09	1.00	0.081

Since the score associated with case 2 is higher than the score associated with case 1, we believe that case 2 should be selected in these circumstances. Although its index variation is lower, there is greater confidence that its solution will give good results.

## 5 Conclusions and Future Directions

A case base can be compared to human memory. It makes it possible to retrieve experiences that were successful in the past and apply them to identical situations. However, solutions that have been previously successful do not always lead to good results in new problems. In this work, we aimed to develop a methodology for retrieving cases based on similarity and success. The motivation for the topic arose after we found that the most studies focused on case retrieval do not consider the success ratio associated with each solution. The cases in a case base serve as a reference for several new problems. Many of these problems fail, which shows that the reference solution is not 100% successful. For this reason, the success ratio should also be a factor to consider when selecting cases. Another aspect to note is that whenever a quantitative evaluation measure is available, such as a well-being index, it is preferable to select the case that leads, on average, to the best results. Therefore, we concluded that the definition of a methodology for the retrieval phase needs to include additional metrics besides similarity, namely the success ratio and a measure to quantify the results obtained. For future approaches, it would be important to test the proposed methodology in different situations in order to see if it is necessary to assign different weights to each of the key concepts: similarity, index variation and success ratio.

**Acknowledgements** This work has been supported by FCT—Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020, and the PhD grant: 2022.12728.BD.



CIÊNCIA, TECNOLOGIA  
E ENSINO SUPERIOR



## References

1. Thrun, S., Pratt, L.: Learning to Learn. Springer US, Boston, MA (1998). <https://doi.org/10.1007/978-1-4615-5529-2>.
2. Clifton, J.R., Frohnsdorff, G.: Applications of Computers and Information Technology. In: Handbook of Analytical Techniques in Concrete Science and Technology. pp. 765–799. Elsevier (2001). <https://doi.org/10.1016/b978-081551437-4.50021-7>.
3. Poggio, T., Bizzi, E.: Generalization in Vision and Motor Control. *Nature*. 431, 768–774 (2004). <https://doi.org/10.1038/nature03014>.



4. Pal, S.K., Shiu, S.C.K.: *Foundations of Soft Case-Based Reasoning*. John Wiley & Sons, Inc. (2004).
5. Bergmann, R., Althoff, K.-D., Minor, M., Reichle, M., Bach, K.: *Case-Based Reasoning - Introduction and Recent Developments*. *Künstliche Intelligenz*. 9, 5–11 (2009).
6. Soliman, G.A.: Dietary Cholesterol and the Lack of Evidence in Cardiovascular Disease. *Nutrients*. 10, (2018). <https://doi.org/10.3390/nu10060780>.
7. Aamodt, A., Plaza, E.: *Case-based Reasoning: Foundational Issues, Methodological Variations, and System Approaches*. *Artif. Intell. Commun.* 7, 39–59 (1994). <https://doi.org/10.3390/s120811154>.
8. Smyth, B., McClave, P.: Similarity vs. Diversity. In: Aha, D.W. and Watson, I. (eds.) *Proceedings of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development*. pp. 347–361. Springer-Verlag, Berlin, Heidelberg (2001). [https://doi.org/10.1007/3-540-44593-5\\_25](https://doi.org/10.1007/3-540-44593-5_25).
9. Ahn, H., Kim, K.: Using Genetic Algorithms to Optimize Nearest Neighbors for Data Mining. *Ann. Oper. Res.* 163, 5–18 (2008). <https://doi.org/10.1007/s10479-008-0325-2>.
10. Castro, J.L., Navarro, M., Sánchez, J.M., Zurita, J.M.: Loss and gain functions for CBR retrieval. *Inf. Sci. (Ny)*. 179, 1738–1750 (2009). <https://doi.org/10.1016/j.ins.2009.01.017>.
11. Kang, Y.-B., Krishnaswamy, S., Zaslavsky, A.: A Retrieval Strategy for Case-Based Reasoning Using Similarity and Association Knowledge. *IEEE Trans. Cybern.* 44, 473–487 (2014). <https://doi.org/10.1109/TCYB.2013.2257746>.
12. Aljuboori, A., Meziane, F., Parsons, D.: A New Strategy for Case-Based Reasoning Retrieval Using Classification Based on Association. In: Perner, P. (ed.) *Machine Learning and Data Mining in Pattern Recognition*. pp. 326–340. Springer International Publishing (2016). [https://doi.org/10.1007/978-3-319-41920-6\\_24](https://doi.org/10.1007/978-3-319-41920-6_24).
13. Fei, L., Feng, Y.: A Novel Retrieval Strategy for Case-Based Reasoning Based on Attitudinal Choquet Integral. *Eng. Appl. Artif. Intell.* 94, (2020). <https://doi.org/10.1016/j.engappai.2020.103791>.
14. QRISK2, <https://qrisk.org/2017/>, last accessed 2022/10/26.
15. Davidson, J.A., Banerjee, A., Smeeth, L., McDonald, H.I., Grint, D., Herrett, E., Forbes, H., Pebody, R., Warren-Gash, C.: Risk of Acute Respiratory Infection and Acute Cardiovascular Events Following Acute Respiratory Infection Among Adults With Increased Cardiovascular Risk in England Between 2008 and 2018: a Retrospective, Population-Based Cohort Study. *Lancet Digit. Heal.* 3, e773–e783 (2021). [https://doi.org/10.1016/S2589-7500\(21\)00203-X](https://doi.org/10.1016/S2589-7500(21)00203-X).

# **Machine Learning and Deep Learning**

# Forming of Validation Dataset for Deep Learning Based Model of Medical Image Grouping



Robert Baždarić, Franko Hržić, Mateja Napravnik, and Ivan Štajduhar

**Abstract** This paper presents a human-machine method for creating multimodal, hierarchically organised medical image validation datasets for the purposes of an AI image annotation algorithm. The multimodal image datasets provided here are independent of the training and testing datasets, as they are obtained in different ways from publicly available medical repositories. Objectivity in the formation of the datasets is maintained based solely on DICOM metadata. These are mainly provided by the medical instruments themselves and by computer-assisted parsing of the original metadata. To overcome the inconsistency of metadata from repository to repository, the algorithm incorporates human observation and monitoring that can guarantee higher confidence in morphological uniformity. The presented methodology provides and discusses the use of image parameters for broader and quantity-based selection and grouping of medical images. The rendered groups contain images with the obtained information leading to the general source of the data for easier use and understanding.

**Keywords** DICOM based grouping · Medical image morphology · Medical image grouping

---

R. Baždarić (✉) · F. Hržić · M. Napravnik · I. Štajduhar  
Department of Computer Engineering, University of Rijeka Faculty of Engineering,  
Vukovarska 58, 51000, Rijeka, Croatia  
e-mail: [rbazdaric@riteh.hr](mailto:rbazdaric@riteh.hr)

F. Hržić  
e-mail: [fhrcic@riteh.hr](mailto:fhrcic@riteh.hr)

M. Napravnik  
e-mail: [mnapravnik@riteh.hr](mailto:mnapravnik@riteh.hr)

I. Štajduhar  
e-mail: [istajduh@riteh.hr](mailto:istajduh@riteh.hr)

F. Hržić · I. Štajduhar  
Center for Artificial Intelligence and Cybersecurity, University of Rijeka, Trg braće Mažuranića  
10, 51000, Rijeka, Croatia

## 1 Introduction

The fundamental problem in developing an artificial intelligence (AI) model is certainly the qualitative selection of the learning database. Since the overall database should be divided into several data portions for different learning tasks, in this paper we present the work involved in forming a validation portion of the complex modelling in medical image grouping. Given the enormous prevalence of machine learning (ML) and deep learning (DL) methods in solving a wide range of tasks encountered in normal human life, especially in image-based decision making, it is inevitable that these methods will also be used in medical image analysis. The neural network (NN) is becoming a fundamental building block in the construction of decision support systems (DSS) based on computer-aided diagnosis (CAD). Comparing medicine and biomedicine with other disciplines and professions, the overall contribution of society to development or its influence is certainly weaker. The reason for this is in the need for specific knowledge, but also in the ethical constraints associated with the availability of data. This work is part of a larger project that addresses the problem of building a suitable and general image database for the development of various methods in CAD. Comparable to the ImageNET project, presented in [1], our general task is also to create hierarchically organised medical images under the title “Machine Learning for Knowledge Transfer in Medical Radiology—RadiologyNET” [8]. This work contributes to the validation part of the task to build the AI model for grouping the clinical radiology datasets. The future model in [8] will be based on datasets selected under different aspects from picture archiving and communication system (PACS) of the Clinical Hospital Centre, Rijeka, Croatia. For the quality-based validation process, the work in this article provides the targeted ground-truth datasets that are independently sourced from the publicly available databases. By using the mathematical formalism in terms and applying it to medical image analysis, the task of providing CAD can be divided into a quantitative and a qualitative process. The quantitative process is recognised in the formation of a huge database [8], which will be a prerequisite for the derivation of the qualitative AI model. In the proposal [8], the final grouping will be a synergy of DICOM-based, narrative diagnosis, and last but not least, image-based classification. In the past, there have been similar attempts to create and test expert annotated datasets [2, 3, 6]. From these attempts, the idea of providing different multimodal annotated datasets evolved [7]. Our approach follows this idea, but differs from it by focusing on the objectivity of the image data during the process of image generation, rather than relying on multiple transformations of the images on the way to the final AI user. The problem raised is slowly gaining attention [10], and the efforts of numerous academies and the support of the governments of various countries are moving toward a solution in publicly available datasets. In this work, we search for the validation datasets. The search for the validation datasets is no less challenging than the search for the training and testing datasets. For this purpose, we have used several publicly available sources. They were thoroughly reviewed to assess their compatibility with the task defined above. The sources listed in Table 1 are considered as candidates for the formation of validation datasets. Since one of

our main grouping objectives is to reveal the ground-truth of the images and their morphological similarity, we selected the “National Biomedical Imaging Archive (NBIA) US” or more generally “The Cancer Imaging Archive” (TCIA) in Table 1 for greater consideration. The labelled validation set must be a medically characterised and group-oriented database. It must follow the monohierarchical, multi-axial [3] strategy in annotation, which is the basic idea in the standardisation of the final multimodal datasets. Therefore, the crucial characteristics in the selection of the database sources are the diversity of the human organs examined, the diversity of the modality of examination, the number of images available, and finally the number of subjects examined. The hierarchical order of the characteristics corresponds to their priority in the selection process. It is difficult to endorse a lack of standardisation in the provision and storage of medical images, despite the existence of the solid DICOM standard, which is regularly evolving. The promising attempts have emerged in the last two decades with the standardisation of IRMA [3]. For the purposes of our work, the wide availability of IRMA annotated images would be beneficial. As far as authors are aware, there are no multimodal publicly available annotated data sources, and the work presented in this paper will provide them for the validation portion of the AI learning in [8]. The groups provided are developed strictly based on the image associated data stored as a result of the DICOM standard during the patient examination process. Although in [3] the reasons for the evolving standardisation relies on the medical device conformance, we strongly believe that after a decade of progress in the medical device technology, this fact should be neglected. If there is a non-conformity with standardisation, it is still the uniformly followed in the image formation, which is easy to capture in various computerised data analysis. The main problem we have identified lies in the human factor and the data, which is mainly influenced by human input. Since in this work we will provide grouping based purely on available DICOM metadata, the selection of metadata is carefully worked out and observed during the grouping process. Section 2 then presents the general challenges of grouping or annotation in the validation dataset, using the TCIA dataset as an example. Section 3 presents the methodology to overcome the problem described in Sect. 2. Section 4 presents DICOM-based grouping algorithms that support the methodology explained in Sect. 3. Section 5 discusses the results of the performed grouping of the validation dataset. The paper is concluded with a brief conclusion in Sect. 6.

## 2 Challenges in Selection of the Valuable Source of Medical Images

As noted above, consistent with our intent to strive for objectivity in the storage of medical images, we will closely examine the NBIA repository. The NBIA has a large number of repositories that are preserved through their structure and are accessible through jointly developed indexing and links to sources. It is also enriched by control-

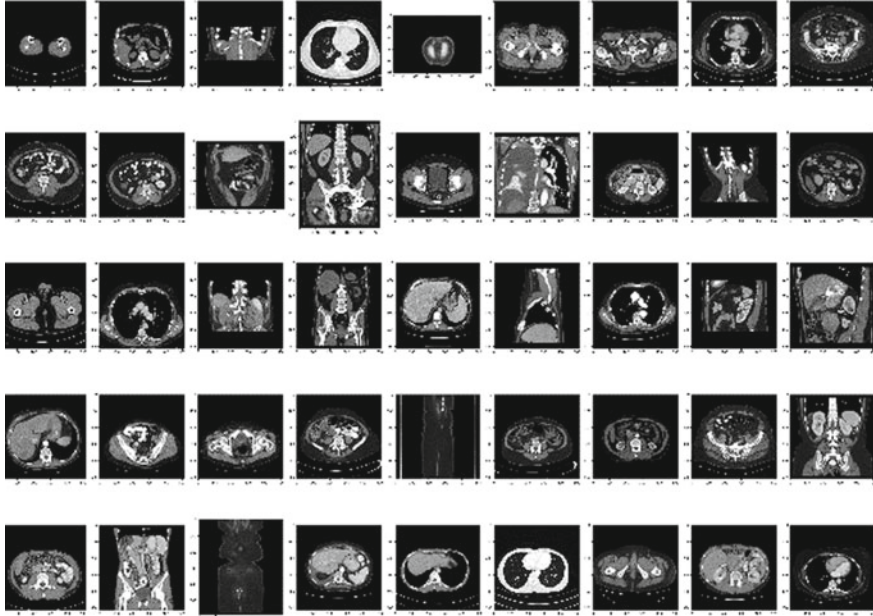
**Table 1** Publicly available repositories

Source	Link	Source	Link
1. Mendely Data	<a href="https://data.mendeley.com/datasets/rscbjbr9sj/2">https://data.mendeley.com/datasets/rscbjbr9sj/2</a>	5. Stanford Institute, MRNet	<a href="https://stanfordmlgroup.github.io/competitions/mrnet/">https://stanfordmlgroup.github.io/competitions/mrnet/</a>
2. National Institute of Health CC, US	<a href="https://nihcc.app.box.com/v/ChestXray-NIHCC">https://nihcc.app.box.com/v/ChestXray-NIHCC</a>	6. Stanford Institute, Mura	<a href="https://stanfordmlgroup.github.io/competitions/mura/">https://stanfordmlgroup.github.io/competitions/mura/</a>
3. National Biomedical Imaging Archive US (TCIA), Cancer	<a href="https://www.cancerimagingarchive.net/collections/">https://www.cancerimagingarchive.net/collections/</a>	7. Medical Segmentation Decathlon	<a href="http://medicaldecathlon.com/">http://medicaldecathlon.com/</a>
4. Stanford Institute, CheXpert	<a href="https://stanfordmlgroup.github.io/competitions/chexpert/">https://stanfordmlgroup.github.io/competitions/chexpert/</a>	8. IRCAD - Hôpitaux Universitaires - 1, FRANCE	<a href="https://www.ircad.fr/research/data-sets/">https://www.ircad.fr/research/data-sets/</a>

lable downloads through the NBIA software. Table 2 shows only a limited selection from the collective and indexed list of available repositories. The NBIA web offers the possibility to search for suitable databases and to make a simple selection of these repositories, adapted to the task-related requirements. As expected, and due to the wide range of DICOM data in the files, the repositories we selected in Table 2 are downloaded in conjunction with the \*.csv manifest, which does not contain all the required image attributes. If we recall the standardisation of IRMA [3], our publicly available databases are formed so that group D is at the top of the hierarchy of image groups. The repositories compiled in Table 2 are strictly differentiated by medical diagnosis and fall into our group of quality-oriented ML databases. This is reminiscent of our earlier distinction between quantitative and qualitative image ML applications, and the emerging requirements for selectivity and hierarchical availability of databases. Means, the targeted database should start from the overall organ based grouped images that fall in hierarchy towards the biological system examination similarity [3]. For example, as stated, the databases that are used and extracted from TCIA are mostly qualitatively oriented, preselected and filtered according to the type of cancer. If we take, e.g. the Cancer Genome Atlas (TCGA) group of repositories that were predominantly selected as data source in our work, TCIA contains only radiological data or DICOM-based stored images, but it comprehensively contains our selection of image attributes, which are explained below. Even in cases where attributes are present, this is not a sufficient condition for automatic labelling, which is our task in providing the quantitative image characteristics. Figure 1 shows a visualisation example based on the TCGA\_LIHC liver CT database.

**Table 2** TCIA selected repositories

Collection	Location	Collection	Location	Collection	Location	Collection	Location
Ultrasound data of a variety of liver masses (B-mode-and-CEUS-Liver)	Liver	ACRIN-FLT-Breast (ACRIN 6688)	Breast	TCGA-THCA	Thyroid	TCGA-LIHC	Liver
Duke-Breast-Cancer-MRI	Breast	Pelvic-Reference-Data	Pelvis, prostate, anus	TCGA-CESC	Cervix	TCGA-KIRC	Kidney
CMMD	Breast	TCGA-BLCA	Bladder	TCGA-ESCA	Esophagus	SPIE-AAAPM Lung CT Challenge	Lung
CPTAC-PDA	Pancreas	PROSTATEx	Prostate	TCGA-STAD	Stomach	CT Colonography (ACRIN 6664)	Colon
CPTAC-UCEC	Uterus	CT Lymph Nodes	Abdomen, mediastinum	TCGA-OV	Ovary	Stage-II-Colorectal-CT	Abdominal, rectal



**Fig. 1** Example of randomly selected images from the TCGA\_LIHC repository, CT examinations

```
data_frame = ({'path_no': [], 'path': [], 'file': [], 'CrossSection': [], 'ImageType': [], 'BodyPartExamined': [], 'Modality': [],
  'PatientPosition': [], 'PatientOrientation': [], 'ContrastBolusAgent': [], 'WindowCenter': [], 'WindowWidth': [],
  'PhotometricInterpretation': [], 'NumberOfSlices': [], 'SliceLocation': [], 'SliceThickness': [], 'PixelSpacing': [],
  'MultienergyCTAcquisition': [], 'ScanOptions': [], 'ImageLaterality': [], 'BreastImplantPresent': [],
  'PositionerPrimaryAngleDirection': [], 'PartialView': [], 'MRAcquisitionType': [], 'AnatomicalOrientationType': [],
  'ScanningSequence': [], 'SequenceVariant': [], 'ScanOptions': [], 'AngioFlag': [], 'SeriesDescription': [],
  'StudyDescription': [], 'SpatialLocationsPreserved': [], 'PatientOrientation': [], 'ImagePositionPatient': [],
  'ImageOrientationPatient': [], 'ExposureTime': [], 'Exposure': []})
```

**Fig. 2** Data selection for examination of repositories

In the following, we present a preliminary way to overcome the problem and form more uniform groups of images that contain the unique attributes not present in the original grouping of Table 2.

### 3 Forming of Distinctive Groups, Constrained Combinations

Our selection of metadata must be distinguishable primarily in terms of the organs examined and the examination modality, but will certainly be sufficient later to form a group of data specific to the medical-diagnostic examinations performed on the subjects. Because our grouping goes for the validation data, and in order to be able



to comment on the final modelling results (possibly fine-tuning), the group labelling must be as accurate as possible. In the absence of specific medical knowledge that extends to instrumentation and diagnosis, our grouping is based on a purely technical approach. This means that our grouping is based on the DICOM metadata present in the images, which includes the setting of the examiner and the formation of the numerical image matrix. Thus, the purely technical information must be closely and directly related to the final goal of evaluating the classes derived from the model, which is yet to be determined and is the future product of unsupervised learning. By using AI and combining supervised and unsupervised learning [4, 5, 9], our task will most likely lead to a medical diagnosis in the near future, characterised by a possibly optimised grouping. Since grouping is a purely human perception, but also the defining parameter in the AI learning process, human involvement and supervision will still be inevitable. The metadata available is far from uniformly stored and is rather dependent on individual examiners, in addition to the original device settings and the different approaches of the device manufacturers in storing DICOM-standardised data. Although DICOM-standardised data is of great use in our task, which is an objective mixture of human and computerised grouping process. The final task of the AI model is to further develop the DICOM standardised data and provide unique representatives that lead to the characteristics of the group as a product of medical diagnosis. Figure 2 shows the Python-based dictionary definition of the data containing the image metadata that forms the basis for our human-supervised and computer-assisted process of objective selection. Our procedure for forming the validation data should not affect the data structure of the repository downloaded from the public sources mentioned above. In this way we have the possibility to trace the problem back to the source of the image and comment objectively on the results. The extracted image in our group contains information about the source path (“path”, Fig. 2), its original filename (“file”, Fig. 2) and is additionally indexed by the path index (“path\_no”, Fig. 2) to avoid complexity in labelling images. With the exception of the “CrossSection”, the rest of the targeted data frame, which is later used for grouping, is associated with DICOM standard image attributes. The selection of DICOM tags should be sufficient for rough grouping of human organ examination, and the technical characteristics of image generation should be sufficient for the particular examination process. Our dendrogram of grouping is shown in Fig. 3. Of the repositories selected in Table 2, only CPTAC-UCEC has the DICOM Body Part Examined attribute unassigned. In this particular repository, we used the “Series Description” and “Study Description” attributes as the basis for the final determination of which organ examination the examiner focused on. The “Modality” and “Patient Position” attributes (attributes to images, stored DICOM-based) are generally present in all TCIA repositories we searched. This is not the case for “Patient Orientation”. For this reason, we chose the “Image Orientation Patient” (IOP) as the source of information to determine the final cross-section of the stored image with respect to the fixed position of the patient or the geometry of the examination instrument. The mentioned attribute consists of a tuple of 6 real numbers in the range  $[-1, 1]$  representing the rotation parameter of the original medically defined three-dimensional examination space of

the body and formed as follows:  $x$  (from right to left),  $y$  (from anterior to posterior) and  $z$  (from inferior to superior). The 6-tuple of information is a mathematical 6-dimensional vector formed from two 3D normalised vectors of projections of the original medical  $x$  and  $y$  axes of the human body:  $IOP = [X_x, X_y, X_z : Y_x, Y_y, Y_z]$ . In the image, we call the resulting and stored plane a cross section (CS). In this particular image grouping task, we are more interested in the coarse distinction between images, i.e., the quantitative distinction. Therefore, we reduce the IOP to four characteristic subsets of CSs, namely: axial, sagittal, coronal and oblique. In Fig. 3, the subgroups mentioned in the last level of the grouping, shorter CS, are planar sections through the space of the human body, orthogonal, but with one exception assigned by Oblique. Variants of the pixel directions of the sections are omitted. As mentioned earlier, the “Oblique” section is a set consisting of all possible sections that are not orthogonal to the original medical geometry of the human body. This approach serves to reduce the number of possible different CSs and their subgroups, which we consider a more qualitative approach to our grouping and project perspective in the later stages. If the CS is purely orthogonal to the original human geometry or the orthogonality error is infinitesimally small (allowing for calibration offsets), the image is automatically classified as a member of one of three distinguishable and accurate sets. The additional attribute for the set is provided by the “CrossSection” parameter (Fig. 2), namely “Axial”, “Sagittal” and “Coronal”. The attribute IOP, which contains the information CS, when purely orthogonal, consists of only two non-zero values. These values are the position of the transformed pixel plane, which is oriented to the position of our original human CS plane  $x$ - $y$ , as explained earlier. The recognised combinations are divided into three precisely distinguishable sets as follows:  $Axial = \{[1, 0, 0, 0, 1, 0], [0, 1, 0, 1, 0, 0]\}$ ,  $Sagittal = \{[0, 1, 0, 0, 0, 1], [0, 0, 1, 0, 1, 0]\}$ ,  $Coronal = \{[1, 0, 0, 0, 0, 1], [0, 0, 1, 1, 0, 0]\}$ . All CSs that fall out of purely orthogonal groups form the attribute “Oblique” for the respective group:  $Oblique = \{\forall IOP : IOP \notin Axial \wedge IOP \notin Sagittal \wedge IOP \notin Coronal\}$ . This group is very sensitive as it contains images that undoubtedly correspond to a specific setup in medical diagnostics of certain organ examinations, but may also be a specific result of AI routines in instruments that are rapidly increasing. Accordingly, this group will receive a special attention in our future model validation.

## 4 Process of Grouping and Algorithms

The process of grouping validation images, using the human-machine approach to grouping is shown in Fig. 4. Humans are involved in the selection of sources and later in the formation of groups of interest. First, in source selection, human participation means selecting a repository that can provide a sufficient number of examinations of specific organs on a sufficient number of subjects.

Downloading the selection must be done in accordance with the requirements and conditions (registration, NBIA software, etc.) specified and regulated by the source.

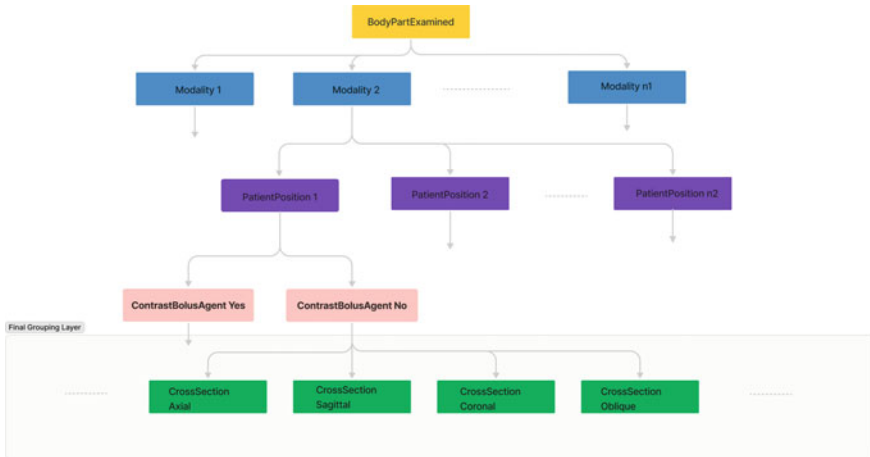


Fig. 3 Dendrogram of image grouping

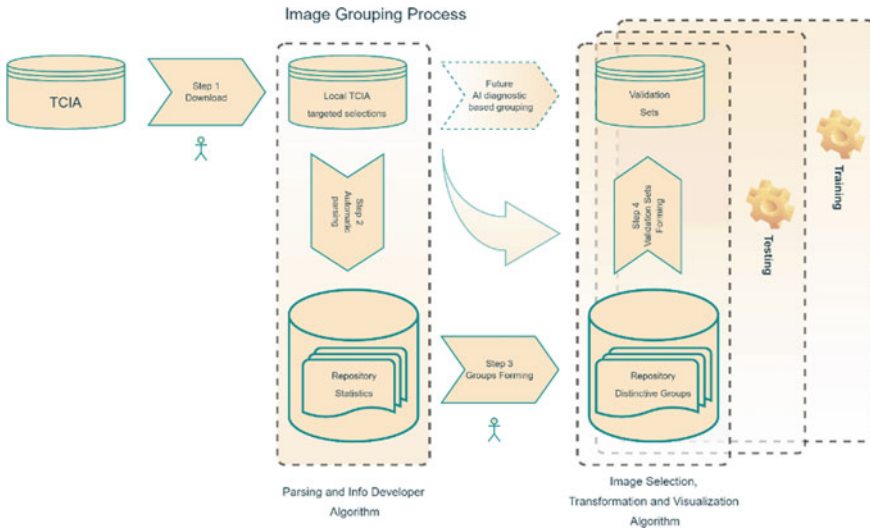
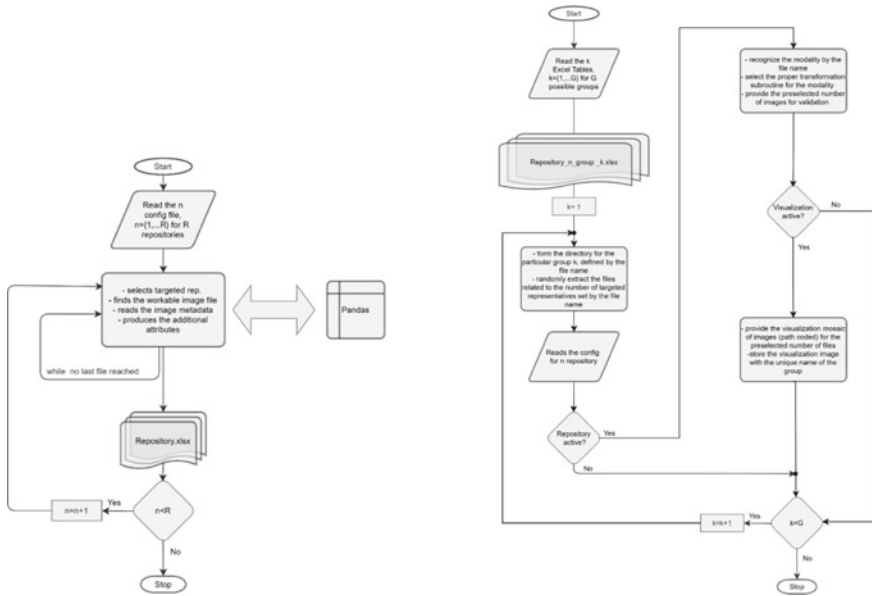


Fig. 4 Human-machine process of medical images' grouping

After downloading the selection from the source (Fig.4, TCIA), the downloaded repository is analysed using the algorithm in Fig.5a. Figure 2 shows the specific and task-related data that the algorithm will read or compute, as explained in Sect.3. Second, after the parsing with the computer algorithm has provided information tables for each repository, one will form the respective data group by simply selecting the same attributes, starting with the coarse information in the hierarchy (organ, modality) and ending with the specific and exact attributes related to the pixel matrix information (Fig.3). The filtered Excel tables are stored in the temporary directory



**Fig. 5** Algorithms: **a** Parsing and info developer algorithm, **b** Image selection, transformation and visualisation algorithm

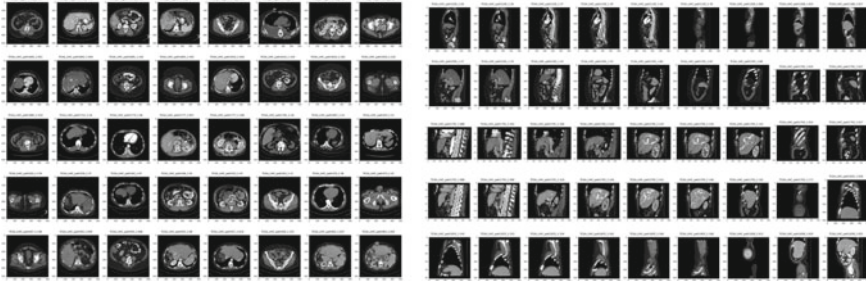
for data extraction. The name of the filtered Excel table contains the identifying information including some parameters

*CT.COLONOGRAPHY\$Colon\_CT\_FFDR\_Contrast\_Axial\_01000.xlsx*

In the grey colour we have the original repository tag, after the sign \$ follows the group selection hierarchy, which is also the unique tag of the group. Of all the possibilities contained in the Excel file, the number in red is the number of randomly selected representatives. The algorithm in Fig. 5b provides a computer routine that extracts the files from the repositories into the newly formed group directories. The algorithm calls the user’s configuration file, which adds to the modality information in the file name a list of commands that select the specific computer subroutines. These subroutines perform the image transformations and visualisation of the groups when specifically selected. As an example, the name of Excel file above shows in blue the modality information CT, which selects a specific image transformation, starting with the Hounsfield unit transformation and followed by the organ window transformation.

## 5 Results

The process explained in Sect. 4 and the tools explained in Sect. 3 lead to a more homogeneous distinction and labelling of image groups. Our original groups, shown in Fig. 1, which are directly downloaded and considered in the coarser grouping



**Fig. 6** Results from the process presented by the Fig.4 among the TCGA\_LIHC repository shows: **a** TCGA\_LIHC\_Liver\_CT\_FFS\_Contrast\_Axial group of images, **b** TCGA\_LIHC\_Liver\_CT\_FFS\_Contrast\_Sagittal group

of the NBIA (see Table 2), have transformed into several DICOM-based differentiated subgroups. The number of subgroups depends on the maximum number of different attributes found in the repositories. For example, Fig. 6 shows two groups with different representatives that have evolved into more monolithic groups by content compared to Fig. 1. In Fig. 6 only two selected subgroups are shown for clarity. From the repositories selected in Table 2, we distinguished 275 groups of images, all of which evolved through by the grouping strategy in Fig. 3 and the availability of images in the repository. The number of representatives in the group varies. Although we aimed for a number of 1,000 (parameter in the group file name), some of the groups contain a dozen or fewer images, while other groups have more than 250 representatives. Figure 7a shows the number of subgroups distributed among the selected repositories from Table 2. As mentioned earlier, we have focused as much as possible on those repositories that offer a large number of subgroups. In Fig. 7b we see how many images we were able to retrieve from the repositories for all differentiated 274 subgroups. The list of subgroups can be found in the Appendix. To better illustrate the diversity of groups, Fig. 8 shows various statistics on the number of subgroups if we want to rearrange the grouping hierarchy at a different level of attributes in Fig. 3.

## 6 Conclusion

This work contains 275 labelled groups of medical images. The images are uniformly grouped according to their DICOM metadata, most of which comes from the original medical imaging process. The images are named with a unique name that leads to a common data source that may be useful for understanding future AI classification and prediction results. The grouping is classified as quantitative grouping because the task is geared toward large collections of morphologically uniform images rather than a task-oriented database for diagnostic purposes. The groups are provided with

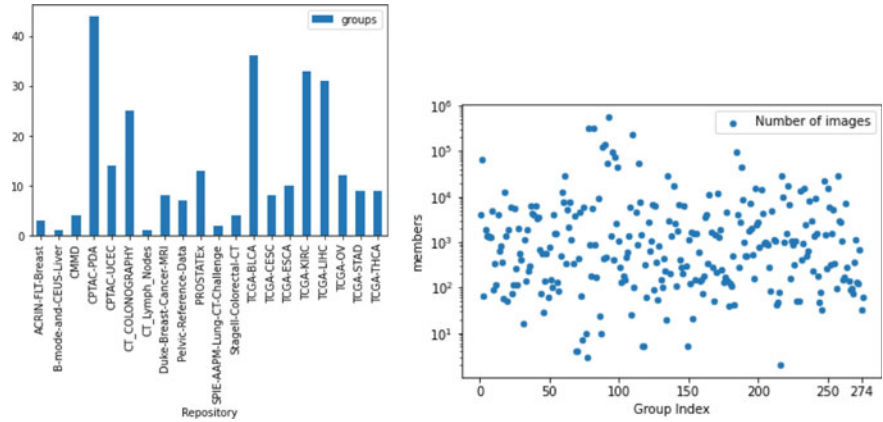


Fig. 7 a Number of subgroups extracted from each repository. b Number of images included in each differentiated subgroup

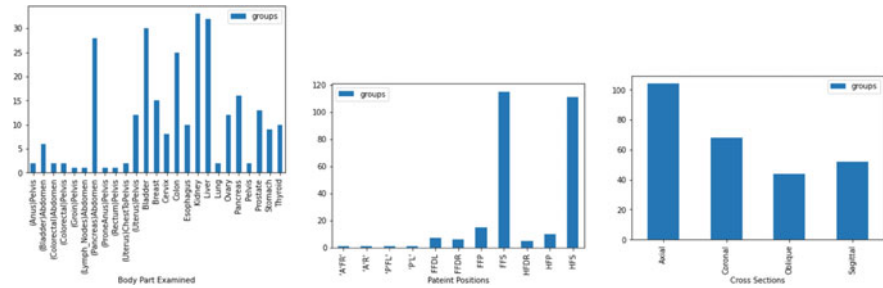


Fig. 8 Number of subgroups displayed at different levels of the grouping hierarchy: a Number of groups for the same “Body Part Examined”, b Number of groups for the same “Patient Position” selected, c Number of groups for the same “Cross Section” selected

their statistical image-related measures. Future work is in the direction of providing automatic web access to the groups, which will also be confirmed by the knowledge of experts before publication. The groups will be linked to the image-based group statistics and the original metadata that can help in annotating the results based on the use of the group.

### A Table of Extracted Groups

See Table 3.

**Table 3**

No.	Origin	Group name	MBRS	No.	Origin	Group name	MBRS	No.	Origin	Group name	MBRS	No.	Origin	Group name	MBRS
1	ACRIN-FLT-Breast	Breast_CT_FFS_Non Contrast_Axial	4096	93	Duke-Breast-Cancer-MRI	Breast_MR_FFP_Contrast_Axial	557034	185	TCGA-KIRC	Kidney_CT_FFS_Contrast_Coronal	3952				
2	ACRIN-FLT-Breast	Breast_CT_HFS_Non Contrast_Axial	64128	94	Duke-Breast-Cancer-MRI	Breast_MR_FFP_Contrast_Oblique	1365	186	TCGA-KIRC	Kidney_CT_FFS_Contrast_Oblique	288				
3	ACRIN-FLT-Breast	Breast_CT_HFS_NonContrast_Coronal	65	95	Duke-Breast-Cancer-MRI	Breast_MR_FFP_Non Contrast_Axial	93875	187	TCGA-KIRC	Kidney_CT_FFS_Contrast_Sagittal	2712				
4	B-mode-and-CEUS-Liver	Liver_US	1859	96	Duke-Breast-Cancer-MRI	Breast_MR_FFP_Non Contrast_Oblique	407	188	TCGA-KIRC	Kidney_CT_FFS_Non Contrast_Axial	44164				
5	CMMD	Breast_MG_ 'A'FR'	1341	97	Duke-Breast-Cancer-MRI	Breast_MR_HFP_Contrast_Axial	72565	189	TCGA-KIRC	Kidney_CT_FFS_Non Contrast_Coronal	8348				
6	CMMD	Breast_MG_ 'A'R'	1341	98	Duke-Breast-Cancer-MRI	Breast_MR_HFP_Contrast_Oblique	2624	190	TCGA-KIRC	Kidney_CT_FFS_NonContrast_Oblique	986				
7	CMMD	Breast_MG_ 'P'FL'	1260	99	Duke-Breast-Cancer-MRI	Breast_MR_HFP_Non Contrast_Axial	43674	191	TCGA-KIRC	Kidney_CT_FFS_Non Contrast_Sagittal	490				
8	CMMD	Breast_MG_ 'P'L'	1260	100	Duke-Breast-Cancer-MRI	Breast_MR_HFP_Non Contrast_Oblique	1042	192	TCGA-KIRC	Kidney_CT_HFP_Contrast_Axial	492				
9	CPTAC-PDA	(Pancreas)Abdomen_CT_FFS_Contrast_Axial	4724	101	Pelvic-Reference-Data	(Anus)Pelvis_CT_HFP_NonContrast_Axial	274	193	TCGA-KIRC	Kidney_CT_HFS_Contrast_Axial	1777				
10	CPTAC-PDA	(Pancreas)Abdomen_CT_FFS_Contrast_Coronal	87	102	Pelvic-Reference-Data	(Anus)Pelvis_CT_HFS_Non Contrast_Axial	131	194	TCGA-KIRC	Kidney_CT_HFS_Non Contrast_Axial	6973				
11	CPTAC-PDA	(Pancreas)Abdomen_CT_FFS_Contrast_Oblique	345	103	Pelvic-Reference-Data	(Groin)Pelvis_CT_HFS_NonContrast_Axial	178	195	TCGA-KIRC	Kidney_CT_HFS_Non Contrast_Coronal	2034				
12	CPTAC-PDA	(Pancreas)Abdomen_CT_FFS_Contrast_Sagittal	115	104	Pelvic-Reference-Data	(ProneAnus)Pelvis_CT_HFP_Non Contrast_Axial	153	196	TCGA-KIRC	Kidney_CT_HFS_Non Contrast_Oblique	518				
13	CPTAC-PDA	(Pancreas)Abdomen_CT_FFS_NonContrast_Axial	3907	105	Pelvic-Reference-Data	(Rectum)Pelvis_CT_HFS_NonContrast_Axial	458	197	TCGA-KIRC	Kidney_CT_HFS_Non Contrast_Sagittal	157				
14	CPTAC-PDA	(Pancreas)Abdomen_CT_FFS_NonContrast_Coronal	648	106	Pelvic-Reference-Data	Pelvis_CT_HFP_NonContrast_Axial	595	198	TCGA-KIRC	Kidney_MR_FFS_Contrast_Axial	15548				
15	CPTAC-PDA	(Pancreas)Abdomen_CT_FFS_NonContrast_Sagittal	810	107	Pelvic-Reference-Data	Pelvis_CT_HFS_Non Contrast_Axial	1350	199	TCGA-KIRC	Kidney_MR_FFS_Contrast_Coronal	2398				

(continued)

**Table 3** (continued)

No.	Origin	Group name	MBRS	No.	Origin	Group name	MBRS	No.	Origin	Group name	MBRS
16	CPTAC-PDA	(Pancreas)Abdomen_CT_HFDR_Contrast_Axial	352	108	PROSTATEx	Prostate_MR_FFS_Contrast_Axial	5953	200	TCGA-KIRC	Kidney_MR_FFS_Contrast_Oblique	2755
17	CPTAC-PDA	(Pancreas)Abdomen_CT_HFDR_NonContrast_Axial	57	109	PROSTATEx	Prostate_MR_FFS_Contrast_Coronal	25	201	TCGA-KIRC	Kidney_MR_FFS_Contrast_Sagittal	771
18	CPTAC-PDA	(Pancreas)Abdomen_CT_HFS_Contrast_Axial	12807	110	PROSTATEx	Prostate_MR_FFS_Contrast_Oblique	232623	202	TCGA-KIRC	Kidney_MR_FFS_NonContrast_Axial	14607
19	CPTAC-PDA	(Pancreas)Abdomen_CT_HFS_Contrast_Coronal	1399	111	PROSTATEx	Prostate_MR_FFS_Contrast_Sagittal	101	203	TCGA-KIRC	Kidney_MR_FFS_NonContrast_Coronal	4158
20	CPTAC-PDA	(Pancreas)Abdomen_CT_HFS_Contrast_Oblique	51	112	PROSTATEx	Prostate_MR_FFS_NonContrast_Axial	4503	204	TCGA-KIRC	Kidney_MR_FFS_NonContrast_Oblique	908
21	CPTAC-PDA	(Pancreas)Abdomen_CT_HFS_Contrast_Sagittal	1860	113	PROSTATEx	Prostate_MR_FFS_NonContrast_Coronal	870	205	TCGA-KIRC	Kidney_MR_FFS_NonContrast_Sagittal	977
22	CPTAC-PDA	(Pancreas)Abdomen_CT_HFS_NonContrast_Axial	5773	114	PROSTATEx	Prostate_MR_FFS_NonContrast_Oblique	55855	206	TCGA-KIRC	Kidney_MR_FFS_Contrast_Axial	920
23	CPTAC-PDA	(Pancreas)Abdomen_CT_HFS_NonContrast_Coronal	75	115	PROSTATEx	Prostate_MR_FFS_NonContrast_Sagittal	7491	207	TCGA-KIRC	Kidney_MR_FFS_Contrast_Coronal	875
24	CPTAC-PDA	(Pancreas)Abdomen_CT_HFS_NonContrast_Sagittal	116	116	PROSTATEx	Prostate_MR_FFS_Contrast_Oblique	1440	208	TCGA-KIRC	Kidney_MR_FFS_Contrast_Oblique	282
25	CPTAC-PDA	(Pancreas)Abdomen_MR_FFS_Contrast_Axial	5347	117	PROSTATEx	Prostate_MR_FFS_NonContrast_Axial	5	209	TCGA-KIRC	Kidney_MR_FFS_Contrast_Sagittal	86
26	CPTAC-PDA	(Pancreas)Abdomen_MR_FFS_Contrast_Coronal	264	118	PROSTATEx	Prostate_MR_FFS_NonContrast_Coronal	5	210	TCGA-KIRC	Kidney_MR_FFS_NonContrast_Axial	4746
27	CPTAC-PDA	(Pancreas)Abdomen_MR_FFS_Contrast_Oblique	116	119	PROSTATEx	Prostate_MR_FFS_NonContrast_Oblique	317	211	TCGA-KIRC	Kidney_MR_FFS_NonContrast_Coronal	1110
28	CPTAC-PDA	(Pancreas)Abdomen_MR_FFS_NonContrast_Axial	1980	120	PROSTATEx	Prostate_MR_FFS_NonContrast_Sagittal	63	212	TCGA-KIRC	Kidney_MR_FFS_NonContrast_Oblique	1086
29	CPTAC-PDA	(Pancreas)Abdomen_MR_FFS_NonContrast_Coronal	550	121	SPIE-AAPM-Lung-CT-Challenge	Lung_CT_FFS_Contrast_Axial	2748	213	TCGA-KIRC	Kidney_MR_FFS_NonContrast_Sagittal	42
30	CPTAC-PDA	(Pancreas)Abdomen_MR_FFS_NonContrast_Oblique	422	122	SPIE-AAPM-Lung-CT-Challenge	Lung_CT_FFS_NonContrast_Axial	657	214	TCGA-LIHC	Liver_CT_FFDL_NonContrast_Axial	48

(continued)



**Table 3** (continued)

No.	Origin	Group name	MBRS	No.	Origin	Group name	MBRS	No.	Origin	Group name	MBRS
31	CPTAC-PDA	(Pancreas)Abdomen_MR_FFS_ NonContrast_Sagittal	16	123	StageII-Colorectal-CT	(Colorectal)Abdomen_CT_ FFS_Contrast_Axial	1082	215	TCGA- LIHC	Liver_CT_FFDL_ NonContrast_Coronal	81
32	CPTAC-PDA	(Pancreas)Abdomen_MR_ FFS_Contrast_Axial	1120	124	StageII-Colorectal-CT	(Colorectal)Abdomen_CT_ FFS_NonContrast_Axial	6521	216	TCGA- LIHC	Liver_CT_FFDL_ NonContrast_Sagittal	2
33	CPTAC-PDA	(Pancreas)Abdomen_MR_ FFS_Contrast_Coronal	143	125	StageII-Colorectal-CT	(Colorectal)Pelvis_ CT_FFS_Contrast_Axial	92	217	TCGA- LIHC	Liver_CT_FFS_ Contrast_Axial	28154
34	CPTAC-PDA	(Pancreas)Abdomen_MR_ FFS_NonContrast_Axial	1824	126	StageII-Colorectal-CT	(Colorectal)Pelvis_CT_FFS_ NonContrast_Axial	6155	218	TCGA- LIHC	Liver_CT_FFS_ Contrast_Coronal	9506
35	CPTAC-PDA	(Pancreas)Abdomen_MR_ FFS_NonContrast_Coronal	230	127	TCGA-BLCA	(Bladder)Abdomen_CT_ FFS_Contrast_Axial	361	219	TCGA- LIHC	Liver_CT_FFS_ Contrast_Oblique	1015
36	CPTAC-PDA	(Pancreas)Abdomen_MR_ FFS_NonContrast_Oblique	306	128	TCGA-BLCA	(Bladder)Abdomen_CT_ FFS_Contrast_Coronal	53	220	TCGA- LIHC	Liver_CT_ FFS_Contrast_Sagittal	1562
37	CPTAC-PDA	(Pancreas)CT_ FFS_Contrast_Axial	6391	129	TCGA-BLCA	(Bladder)Abdomen_CT_ FFS_NonContrast_Axial	819	221	TCGA- LIHC	Liver_CT_FFS_ NonContrast_Axial	16915
38	CPTAC-PDA	(Pancreas)CT_FFS_ Contrast_Coronal	4359	130	TCGA-BLCA	(Bladder)Abdomen_CT_ FFS_NonContrast_Coronal	150	222	TCGA- LIHC	Liver_CT_FFS_ NonContrast_Coronal	6699
39	CPTAC-PDA	(Pancreas)CT_FFS_ Contrast_Sagittal	4126	131	TCGA-BLCA	(Bladder)Abdomen_CT_ FFS_NonContrast_Sagittal	103	223	TCGA- LIHC	Liver_CT_FFS_ NonContrast_Oblique	158
40	CPTAC-PDA	(Pancreas)CT_FFS_ NonContrast_Axial	6239	132	TCGA-BLCA	(Bladder)Abdomen_PT_ FFS_NonContrast_Axial	430	224	TCGA- LIHC	Liver_CT_FFS_Non Contrast_Sagittal	1074
41	CPTAC-PDA	(Pancreas)CT_FFS_ NonContrast_Coronal	3386	133	TCGA-BLCA	(Bladder)CT_ FFP_Contrast_Axial	231	225	TCGA- LIHC	Liver_CT_FFS_ Contrast_Axial	876
42	CPTAC-PDA	(Pancreas)CT_FFS_ NonContrast_Sagittal	3464	134	TCGA-BLCA	(Bladder)CT_FFP_ Contrast_Coronal	20	226	TCGA- LIHC	Liver_CT_FFS_ Contrast_Coronal	309
43	CPTAC-PDA	(Pancreas)MR_ FFS_Contrast_Axial	675	135	TCGA-BLCA	(Bladder)CT_FFS_ Contrast_Axial	28333	227	TCGA- LIHC	Liver_CT_ FFS_NonContrast_Axial	1029
44	CPTAC-PDA	(Pancreas)MR_ FFS_Contrast_Coronal	76	136	TCGA-BLCA	(Bladder)CT_FFS_ Contrast_Coronal	2963	228	TCGA- LIHC	Liver_CT_FFS_ NonContrast_Coronal	52
45	CPTAC-PDA	(Pancreas)MR_FFS_ NonContrast_Axial	579	137	TCGA-BLCA	(Bladder)CT_FFS_ Contrast_Oblique	210	229	TCGA- LIHC	Liver_CT_FFS_ NonContrast_Sagittal	50

(continued)

**Table 3** (continued)

No.	Origin	Group name	MBRs	No.	Origin	Group name	MBRs	No.	Origin	Group name	MBRs
46	CPTAC-PDA	Pancreas_MR_HFS_ NonContrast_Coronal	28	138	TCGA-BLCA	Bladder_CT_HFS_ Contrast_Sagittal	1370	230	TCGA-LIHC	Liver_MR_HFS_Contrast_Axial	13796
47	CPTAC-PDA	Pancreas_MR_HFS_ NonContrast_Oblique	136	139	TCGA-BLCA	Bladder_CT_HFS_ NonContrast_Axial	17239	231	TCGA-LIHC	Liver_MR_HFS_Contrast_Coronal	422
48	CPTAC-PDA	Pancreas_MR_HFS_Contrast_Axial	576	140	TCGA-BLCA	Bladder_CT_HFS_ NonContrast_Coronal	2585	232	TCGA-LIHC	Liver_MR_HFS_Contrast_Oblique	964
49	CPTAC-PDA	Pancreas_MR_HFS_Contrast_Coronal	60	141	TCGA-BLCA	Bladder_CT_HFS_ NonContrast_Oblique	107	233	TCGA-LIHC	Liver_MR_HFS_ NonContrast_Axial	15323
50	CPTAC-PDA	Pancreas_MR_HFS_ NonContrast_Axial	793	142	TCGA-BLCA	Bladder_CT_HFS_ NonContrast_Sagittal	1712	234	TCGA-LIHC	Liver_MR_HFS_ NonContrast_Coronal	2430
51	CPTAC-PDA	Pancreas_MR_HFS_ NonContrast_Coronal	144	143	TCGA-BLCA	Bladder_CT_HFS_Contrast_Axial	6575	235	TCGA-LIHC	Liver_MR_HFS_ NonContrast_Oblique	483
52	CPTAC-PDA	Pancreas_MR_HFS_ NonContrast_Oblique	103	144	TCGA-BLCA	Bladder_CT_HFS_Contrast_Coronal	403	236	TCGA-LIHC	Liver_MR_HFS_ NonContrast_Sagittal	652
53	CPTAC-UCEC	(Uterus/ChestTo)Pelvis_CT_HFS_Contrast_Axial	3998	145	TCGA-BLCA	Bladder_CT_HFS_Contrast_Oblique	274	237	TCGA-LIHC	Liver_MR_HFS_Contrast_Axial	8124
54	CPTAC-UCEC	(Uterus/ChestTo)Pelvis_CT_HFS_ NonContrast_Axial	1527	146	TCGA-BLCA	Bladder_CT_HFS_Contrast_Sagittal	210	238	TCGA-LIHC	Liver_MR_HFS_Contrast_Coronal	2846
55	CPTAC-UCEC	(Uterus)Pelvis_CT_HFS_NonContrast_Axial	1600	147	TCGA-BLCA	Bladder_CT_HFS_ NonContrast_Axial	6206	239	TCGA-LIHC	Liver_MR_HFS_Contrast_Oblique	237
56	CPTAC-UCEC	(Uterus)Pelvis_CT_HFS_NonContrast_Coronal	132	148	TCGA-BLCA	Bladder_CT_HFS_ NonContrast_Coronal	293	240	TCGA-LIHC	Liver_MR_HFS_Contrast_Sagittal	1349
57	CPTAC-UCEC	(Uterus)Pelvis_CT_HFS_NonContrast_Oblique	85	149	TCGA-BLCA	Bladder_CT_HFS_ NonContrast_Sagittal	5	241	TCGA-LIHC	Liver_MR_HFS_ NonContrast_Axial	9240
58	CPTAC-UCEC	(Uterus)Pelvis_CT_HFS_ NonContrast_Sagittal	483	150	TCGA-BLCA	Bladder_MR_HFS_Contrast_Axial	1115	242	TCGA-LIHC	Liver_MR_HFS_ NonContrast_Coronal	1560
59	CPTAC-UCEC	(Uterus)Pelvis_MR_HFS_Contrast_Axial	12291	151	TCGA-BLCA	Bladder_MR_HFS_Contrast_Coronal	122	243	TCGA-LIHC	Liver_MR_HFS_ NonContrast_Oblique	177
60	CPTAC-UCEC	(Uterus)Pelvis_MR_HFS_Contrast_Coronal	7615	152	TCGA-BLCA	Bladder_MR_HFS_Contrast_Axial	1642	244	TCGA-LIHC	Liver_MR_HFS_ NonContrast_Sagittal	46

(continued)

Table 3 (continued)

No.	Origin	Group name	MIBRS	No.	Origin	Group name	MIBRS	No.	Origin	Group name	MIBRS
61	CPTAC- UCEC	(Uterus)Pelvis_ MR_HFS_Contrast_ Oblique	29261	153	TCGA-BLCA	Bladder_ MR_HFS_ Contrast_Sagittal	21	245	TCGA- OV	Ovary_CT_ FFDR_Contrast_Axial	78
62	CPTAC- UCEC	(Uterus)Pelvis_MR_ HFS_Contrast_ Sagittal	347	154	TCGA-BLCA	Bladder_MR_ HFS_Non Contrast_Axial	1596	246	TCGA- OV	Ovary_CT_ FFP_Contrast_ Axial	33
63	CPTAC- UCEC	(Uterus)Pelvis_ MR_HFS_ NonContrast_Axial	5208	155	TCGA-BLCA	Bladder_MR_ HFS_NonContrast_ Coronal	151	247	TCGA- OV	Ovary_CT_ HFS_Contrast_ Axial	21972
64	CPTAC- UCEC	(Uterus)Pelvis_ MR_HFS_NonContrast_ Coronal	2104	156	TCGA-BLCA	Bladder_MR_ HFS_NonContrast_ Oblique	1590	248	TCGA- OV	Ovary_CT_ HFS_Contrast_Coronal	1414
65	CPTAC- UCEC	(Uterus)Pelvis_ MR_HFS_Non Contrast_Oblique	7797	157	TCGA-BLCA	Bladder_MR_HFS_ NonContrast_Sagittal	605	249	TCGA- OV	Ovary_CT_ HFS_Contrast_ Sagittal	263
66	CPTAC- UCEC	(Uterus)Pelvis_ MR_HFS_NonContrast_ Sagittal	1149	158	TCGA-BLCA	Bladder_MR_HFS_ NonContrast_Axial	142	250	TCGA- OV	Ovary_CT_ HFS_NonContrast_ Axial	14951
67	CT_COLONO GRAPHY	Colon_CT_FFDR_ Contrast_Axial	338	159	TCGA-BLCA	Bladder_MR_ HFS_NonContrast_ Coronal	147	251	TCGA- OV	Ovary_CT_ HFS_NonContrast_ Coronal	5941
68	CT_COLONO GRAPHY	Colon_CT_FFDR_ NonContrast_Axial	3888	160	TCGA-BLCA	Bladder_MR_ HFS_NonContrast_ Oblique	74	252	TCGA- OV	Ovary_CT_ HFS_NonContrast_ Sagittal	316
69	CT_COLONO GRAPHY	Colon_CT_ FFDR_NonContrast_ Coronal	4	161	TCGA-BLCA	Bladder_MR_ HFS_NonContrast_ Sagittal	181	253	TCGA- OV	Ovary_CT_ HFS_Contrast_ Axial	1576
70	CT_COLONO GRAPHY	Colon_ CT_FFDR_Non Contrast_Sagittal	4	162	TCGA-BLCA	Bladder_PT_ HFS_NonContrast_ Axial	3468	254	TCGA- OV	Ovary_ CT_HFS_Non Contrast_Axial	5191
71	CT_COLONO GRAPHY	Colon_CT_FFDR_ Contrast_Axial	444	163	TCGA-CESC	Cervix_MR_ HFS_NonContrast_ Axial	4056	255	TCGA- OV	Ovary_CT_HFS_ NonContrast_Coronal	330
72	CT_COLONO GRAPHY	Colon_CT_FFDR_ NonContrast_Axial	4584	164	TCGA-CESC	Cervix_MR_HFS_Non Contrast_Coronal	338	256	TCGA- OV	Ovary_CT_HFS_ NonContrast_Oblique	96
73	CT_COLONO GRAPHY	Colon_CT_FFDR_ NonContrast_Coronal	64	165	TCGA-CESC	Cervix_MR_HFS_Non Contrast_Oblique	11056	257	TCGA- STAD	Stomach_CT_HFS_ NonContrast_Axial	29022
74	CT_COLONO GRAPHY	Colon_CT_FFDR_ NonContrast_Sagittal	7	166	TCGA-CESC	Cervix_MR_HFS_Non Contrast_Sagittal	1685	258	TCGA- STAD	Stomach_CT_HFS_ NonContrast_Coronal	2869
75	CT_COLONO GRAPHY	Colon_CT_FPP_ Contrast_Axial	5894	167	TCGA-CESC	Cervix_MR_HFS_Non Contrast_Axial	286	259	TCGA- STAD	Stomach_CT_HFS_ NonContrast_Sagittal	2820

(continued)



## References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. (2009) 248–255
2. Dimitrovski, I., Kocev, D., Loskovska, S., Džeroski, S.: Hierarchical annotation of medical images. *Pattern Recognition* **44**(10-11) (2011) 2436–2449
3. Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., Wein, B.B.: The irma code for unique classification of medical images. In: *Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*. Volume 5033., SPIE (2003) 440–451
4. Manojlović, T., Ilić, D., Miletić, D., Štajduhar, I.: Using dicom tags for clustering medical radiology images into visually similar groups. In: *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*, Science and Technology Publications (2020) 510–517
5. Manojlović, T., Štajduhar, I.: Deep semi-supervised algorithm for learning cluster-oriented representations of medical images using partially observable dicom tags and images. *Diagnostics* **11**(10) (2021) 1920
6. Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Radhouani, S., Bakke, B., Kahn, C.E., Hersh, W.: Overview of the clef 2009 medical image retrieval track. In: *Workshop of the Cross-Language Evaluation Forum for European Languages*, Springer (2009) 72–84
7. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology objects in context (roco): a multimodal image dataset. In: *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer (2018) 180–189
8. Riteh: Machine learning for knowledge transfer in medical radiology (2019)
9. Štajduhar, I., Manojlović, T., Hržić, F., Napravnik, M., Glavaš, G., Milanič, M., Tschauner, S., Mamula Saračević, M., Miletić, D.: Analysing large repositories of medical images. In: *International Conference on Bioengineering and Biomedical Signal and Image Processing*, Springer (2021) 179–193
10. Willeminck, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L.R., Summers, R.M., Rubin, D.L., Lungren, M.P.: Preparing medical imaging data for machine learning. *Radiology* **295**(1) (2020) 4

# Deep Learning Based Radiomics to Predict Treatment Response Using Multi-datasets



Thibaud Brochet, Jérôme Lapuyade-Lahorgue, Alexandre Huat, Sébastien Thureau, David Pasquier, Isabelle Gardin, Romain Modzelewski, David Gibon, Juliette Thariat, Vincent Grégoire, Pierre Vera, and Su Ruan

**Abstract** In this work, we present a multitask network with multi datasets to assess the relapse of patients with head-neck and lung cancers after a therapy from both scanner images and patient clinical data. A multitask architecture is developed to realize classification of the multi-type of cancers and relapse prediction tasks using clinical data and radiomics features. Medical imaging requires reliable algorithms for analysis and processing, especially regarding diagnosis and outcome prediction. However, in medical domain, only small datasets are available, this is why we propose to combine several small data sets that contain the same type of images and patient information. We also propose to use Havrda-Charvat cross-entropy, which is a generalized cross-entropy with a parameter  $\alpha$ , as loss function for our training step. It tends toward Shannon cross-entropy when said parameter  $\alpha$  is equal to 1. The influence of the variations of the parameter on classification is assessed. The experiments are carried out on a dataset of 580 patients with two cancer datasets (head-neck or lung). The results assess that Havrda-Charvat entropy has slightly better performances in term of prediction accuracy: 64% of correct prediction for Shannon's entropy and at best 69% of correct prediction for Havrda-Charvat for  $\alpha = 0.2$ . The challenge is to find a suitable value of  $\alpha$ .

---

T. Brochet · J. Lapuyade-Lahorgue (✉) · A. Huat · S. Thureau · I. Gardin · R. Modzelewski · P. Vera · S. Ruan  
LITIS, Eq. Quantif, University of Rouen, Rouen, France  
e-mail: [jerome.lapuyade-lahorgue@univ-rouen.fr](mailto:jerome.lapuyade-lahorgue@univ-rouen.fr)

A. Huat · S. Thureau · I. Gardin · R. Modzelewski · P. Vera  
Centre Henri Becquerel, Rouen, France

A. Huat · D. Gibon  
Société Aquilab, Lille, France

D. Pasquier  
Centre Oscar Lambret, Département de radiothérapie, Lille, France

J. Thariat  
CLCC Francois Baclesse, Département de radiothérapie, Caen, France

V. Grégoire  
Département de radiothérapie, Centre Léon Berard, Lyon, France

**Keywords** Deep learning · CT scan · Radiomics · Outcome prediction · Classification

## 1 Introduction

This paper means to study deep neural networks [1] for outcome prediction in both lung and head-neck cancers. In the field of radiomics [2, 3], only small data sets are usually available for outcome prediction. Increasing the datasets and choosing a relevant loss function are primary factors for successful outcome prediction. When categorical prediction is the goal, the loss function turns out to be a cross-entropy derived from an entropy formula. Entropies are a measure of the information contained in random data. Such measures of information are related to disorder in samples of random variable. In [4] are presented several different entropies. In both tasks of classification and prediction, the cross-entropy comes from a specific entropy measure and is tasked to measure the differences between the prediction and reality. In most neural networks used for classification and/or prediction, a Shannon related cross-entropy is the most common and widely used loss function for segmentation [5], classification [6], and detection [7] plus several other uses [8–11]. In [12], several ways of picking an entropy and its corresponding divergence are described. Among them, Shannon entropy is extended by replacing the logarithm with another specific function. Cross-entropies are defined by replacing the counting measure (resp. Lebesgue measure for continuous case) by a Radon-Nykodim derivative between probabilities measures. Shannon's entropy has been declined in other entropies like Renyi [13], Tsallis [14] and Havrda-Charvat [15]. In this article, we study a specific generalization of Shannon's cross-entropy: Havrda-Charvat's cross-entropy [16]. This class of entropies has the particularity that it can be adapted with one parameter  $\alpha$  and we find back Shannon's entropy when the value of the parameter is equal to 1. To increase the size of the dataset, we propose to combine several of the same type of datasets whose difference is the type of cancer observed. A classification task can thus be added. Classification is used to automatically identify what type of cancer the patient experiences [17, 18] or identify relevant outcomes after treatment, like survival expectation [19] or relate to the treatment [2]. Relapse in cancer is a huge concern for physicians [20], as it can dramatically threaten the outcome and life expectancy of patients for the worse. The innovation lies in the performance increase from Shannon to Havrda-Charvat entropies in the context of cancer relapse prediction with Computed Tomography data for patients suffering from head and neck (H&N) and lung cancers, along with clinical data. Moreover, we particularly study the parameter's value to see what its impact on predicting these relapses in both kinds of cancer is. The contributions exposed in this article are: a U-Net based multitask, multi datasets network carrying out reconstruction, classification and segmentation at the same time; the use of Havrda-Charvat entropy to design a novel loss function and the acquisition of good results on both lung and head-neck cancers datasets.

The paper unfolds as follows. First, we explain how categorical Shannon cross-entropy formula is defined and how it can be generalized to Havrda-Charvat and the second part describes our experiments and comparison between both entropies.

## 2 Entropy

As our study regards a prediction, we focus solely on finite-state random variables whose state-space is provided by the counting measure.

### 2.1 Havrda-Charvat Cross-Entropy

There is some ways to generalize Shannon entropy as described in [12]. The Shannon entropy is described as follows:

$$H(q) = - \sum_{i=1}^k h(q_i) \tag{1}$$

where  $h(u) = u \log(u)$ .  $h$  is a convex function that verifies  $h(1) = 0$ . The idea is to pick another function with similar properties. The Havrda-Charvat entropy is defined by picking:

$$h_\alpha(u) = \frac{u^\alpha - u}{\alpha - 1}, \tag{2}$$

where  $\alpha > 0$  and is expressed by:

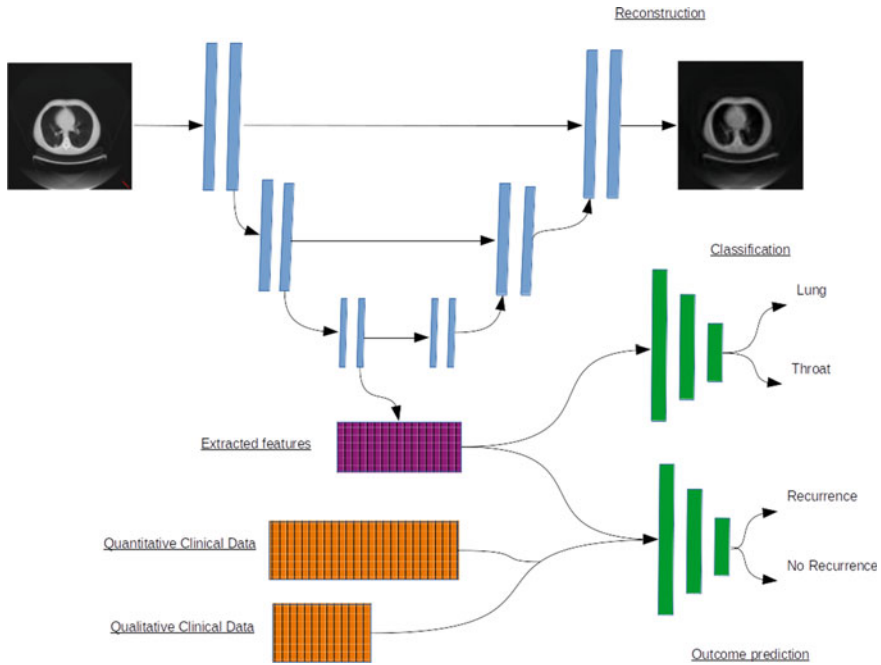
$$H_\alpha(q) = \frac{1}{\alpha - 1} \times \left[ 1 - \sum_{i=1}^k q_i^\alpha \right] \tag{3}$$

The associated cross-entropy is becomes:

$$H_\alpha(q : p) = \frac{1}{\alpha - 1} \times \left[ 1 - \sum_{i=1}^k q_i^{\alpha-1} p_i \right] \tag{4}$$

As for the usual cross-entropy, the Havrda-Charvat cross-entropy forces the prediction  $q$  to be the closest as possible to  $p$  with  $p$  being a Dirac distribution. Indeed, if  $p = \delta_{i_0}$ , then  $H_\alpha(q : p) = \frac{1 - q_{i_0}^{\alpha-1}}{\alpha - 1}$  and the function  $u \rightarrow \frac{1 - u^{\alpha-1}}{\alpha - 1}$  is decreasing and its minimum is reached at 1.





**Fig. 1** Architecture of the network for relapse prediction (T3) with participation from two other tasks (T1: image reconstruction, T2: cancer classification)

### 3 Neural Network Architecture for Relapse Prediction

The proposed network used for relapse prediction is a multitask architecture including a U-Net structure and two branches performing classification and prediction tasks. The architecture is represented in Fig. 1

Three main tasks are taken care of by the architecture. T1 is the reconstruction task, carried out by the U-Net part. It allows to determine if the features extracted by the encoder are relevant for prediction and classification and can be used to reconstruct the whole scan at the same time. The loss function selected for this is the mean squared error. It is presented as follows.

$$L_{rec} = \frac{1}{N} \sum_{n=1}^N \|y_n - \hat{y}_n\|^2, \tag{5}$$

T2 is the classification task. It is used to decide, from the features described earlier and the clinical data, if the cancer is a head-neck cancer or a lung cancer. It works with dense layers and end up in a binary prediction. The loss function used in this branch is Shannon’s binary cross-entropy, as described in Eq. 6 for two classes represented by true classes  $p$  and predicted classes  $q$ ,  $N$  being the sample’s size.

$$L_{\text{classif}} = -\frac{1}{N} \sum_{n=1}^N [p_n \times \log(q_n) + (1 - p_n) \times \log(1 - q_n)] \tag{6}$$

T3 is the carrying out the prediction. It is used to decide, from the images features and table data , if the current patient risks to experience a relapse of their cancer. It is realized with dense layers and end up in a binary prediction. The prediction task’s loss function will be the subject of the tests. We will determine which one gives the most accurate results.

$$L_{\text{pred},\alpha} = \frac{1}{\alpha - 1} \times \left[ 1 - \frac{1}{N} \sum_{i=1}^N (q_n^{\alpha-1} p_n + (1 - q_n)^{\alpha-1} (1 - p_n)) \right] \tag{7}$$

The final loss function of the network is the sum of all three losses.

$$L_{\text{total}} = L_{\text{rec}} + L_{\text{classif}} + L_{\text{pred},\alpha} \tag{8}$$

The prediction task will be what is studied in this article, the other two tasks being used to enhance the prediction’s results by helping the feature extraction.

## 4 Experimentations

### 4.1 Datasets

We use two datasets, one is composed of 434 patients who suffer from head-neck cancer and the other of 146 patients suffering from lung cancer. Scanner images used as inputs in the network have been resized to the following dimensions: (128 × 128 x 64 voxels ). As for patient data used as secondary input in the network, they are of two types: quantitative information and qualitative one, as presented in the following Tables 1 and 2

Our experimentations consist of comparing Havrda-Charvat’s accuracy on both datasets to Shannon’s.

### 4.2 Evaluation Method

As we are currently realizing this study on small datasets, a strategy to validate the results is mandatory. We propose to use the k-fold cross-validation. Cross-validation is in fact a resampling procedure realized to evaluate models when using small data samples. In our work, we use a 5-fold cross-validation.

**Table 1** Quantitative data

Clinical data	Modality
Hemoglobin	g/dL
Lymphocytes	Giga/L
Leucocytes	Giga/L
Thrombocytes	Giga/L
Albumin	g/L
Duration of treatment	Days
Total dose of irradiation	Gy
Number of fractions	/
Average dose per fraction	Gy
Weight both at the start and end of treatment	kg

**Table 2** Qualitative data

Clinical data	Modality
Gender	M/F
Use of tobacco	Smoker, non-smoker, formerly
Use of induction chemotherapy	Yes/no
Use of concomitant chemotherapy	Yes/no
TNM	Tumor, node, metastasis

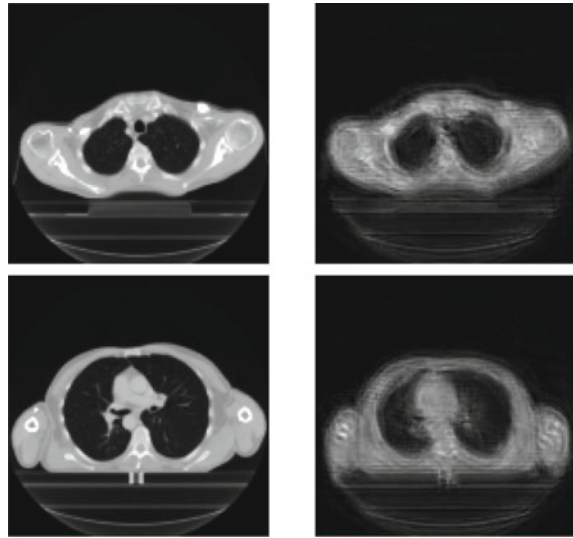
And, as an evaluation metric, accuracy is proposed. It consists, in our study, in comparing the value of the reality and the predicted output, and computing the percentage of correct predictions as our accuracy. To further study the meaning of this value, we also compute the sensitivity and specificity of the prediction. Hereafter are the formula used for all three calculations (eq:Sen,eq:Spe,eq:Acc), with TN being true negative, TP being true positive, FN being false negative and FP being false positive.

$$Sensitivity = \frac{TP}{TP + FN} \quad (9)$$

$$Specificity = \frac{TN}{TN + FP} \quad (10)$$

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of samples}} \quad (11)$$

**Fig. 2** Images : original inputs(left) versus reconstructed slices(right)



### 4.3 Results

The results achieved during the testing phase are presented hereafter. Reconstructed images are used here to assess the relevance of the features extracted by the encoder for the other branches. Therefore, this branch's performance is secondary, because the main goal is the prediction of relapse. The original inputs and reconstructed images are presented in Fig. 2.

The results show that our network can well extract relevant features which are used at same time for the prediction task.

**Havrda-Charvat cross-entropy** Regarding Havrda-Charvat, it was proposed to study the variations of accuracy when its hyperparameter  $\alpha$  varied from 0.1 to 1.3. The achieved p-value assesses if the predictions obtained by Havrda-Charvat are statistically differing from Shannon's with five-fold cross-validation. Two conditions must be met so that we accept the our generalized entropy gives better results than the Shannon entropy. A threshold of 0.05 is taken for the p-value, and the mean result of the 5-fold has to be better to Shannon. Results are described in Table 3.

When we look at the results obtained via our generalized entropy, we can say that, for a lot of the studied values of  $\alpha$ , the final result is inferior to the one obtained via Shannon's loss function. It can be said that the generalized equation can give results that are superior to Shannon's in several cases. However, it is difficult to decide without testing first what  $\alpha$  is relevant for a set application. Its choice remains a complicated, yet here we can demonstrate that Shannon's entropy is not always the best choice.

However, the results achieved with the commonly used Shannon cross-entropy are coherent, judging by the small value of the Standard Deviation. Shannon's cross-

**Table 3** Results obtained by Havrda-Charvat entropy derived loss function in function of  $\alpha$  (p-values inferior to 0.05 are highlighted in bold).

$\alpha$	Accuracy	SD	p-value
0.1	0.62	0.08	0.36
0.2	<b>0.69</b>	0.03	<b>0.01</b>
0.3	0.62	0.03	0.24
0.4	0.63	0.09	0.41
0.5	<b>0.66</b>	0.01	<b>0.02</b>
0.6	0.65	0.06	0.28
0.7	<b>0.67</b>	0.04	<b>0.02</b>
0.8	0.63	0.04	0.31
0.9	0.60	0.04	0.04
1.0	0.64	0.02	N/A (Shannon's entropy)
1.1	0.68	0.05	0.11
1.2	0.63	0.03	0.46
1.3	0.66	0.06	0.23

**Table 4** Results obtained by different tasks combinations with Havrda-Charvat and  $\alpha = 0.2$ .

Tasks	Accuracy	SD	Sensitivity	Specificity
T2, T3	0.64	0.04	0.37	0.79
T1, T3	0.65	0.06	0.18	0.90
T1, T2, T3	0.69	0.04	0.52	0.77

entropy is a stable and reliable formula for loss function, but has difficulties to adapt to non-Riemannian datasets. Results obtained with this entropy has an average accurate rate of 64% from our two datasets. Even if they are lower than Havrda-Charvat's results, they are acceptable.

To demonstrate the advantage of our multi-task network, we propose to assess the importance of each task of the network along with Havrda-Charvat using the most favorable  $\alpha$ . Results are displayed in Table 4.

We notice that results obtained with the combination of all three tasks are the best ones, meaning that the reconstruction and classification tasks can effectively help to find the relevant features for the prediction task.

## 5 Discussion

According to the results, we can conclude that Havrda-Charvat's loss function are equal or superior to Shannon's cross-entropy, depending on the selected  $\alpha$ . It is due to the fact that Havrda-Charvat is a generalized version of Shannon's entropy, able

to perform adequately despite the data's values distribution being non-Riemannian. It is to note that the value of  $\alpha$  is critical for the loss function to perform well, and it needs to be studied to fit the input data, but what  $\alpha$  fits what input? On the other hand, we can conclude that, based on achieved p-values and the corresponding standard deviations, Havrda-Charvat's entropy is less stable than Shannon, as its SD reaches 0.08, where Shannon's is only 0.02. Also, when studying the p-values for several  $\alpha$ , the values achieved via Havrda-Charvat are not statistically significant when compared with Shannon's. The selection of a proper  $\alpha$  is still complicated as it relies heavily on the input data. In perspective, the possibility to automate the selection of the value of this hyperparameter would be very interesting. The goal is to reach a proper area of  $\alpha$  where the loss function gives stable and superior results. Further analysis need to be carried out to assess the input images and the selection of the best  $\alpha$  area, to determine what kind of feature, what kind of neuronal path leads to a zone being the most accurate.

## 6 Conclusion

In this paper, we concluded that, for our data, Havrda-Charvat's formula is giving superior results when compared with Shannon's loss function. Havrda-Charvat best performance on average is 69% of correct relapse prediction while Shannon's is 64%. In medical applications, even a 1 or 2% improvement is interesting. We also show that the combination of several data sets is a good choice in the case of lack of a large data set. A future work would be experimenting on an algorithm to automate the determining of the best value of  $\alpha$  for any input data.

## References

1. Q. Wang, Y. Ma, K. Zhao, and et al., "A Comprehensive Survey of Loss Functions in Machine Learning," *Annals of Data Science*, 2020.
2. A. Amyar, S. Ruan, I. Gardin, and et al., "3D RPET-NET: Development of a 3D PET Imaging Convolutional Neural Network for Radiomics Analysis and Outcome Prediction," *IEEE Trans. on Radiations and Plasma Medical Sciences*, vol. 3, no. 2, pp. 225–231, 2019.
3. Jian Wu, Chunfeng Lian, Su Ruan, Thomas R. Mazur, Sasa Mutic, Mark A. Anastasio, Perry W. Grigsby, Pierre Vera, and Hua Li, "Treatment outcome prediction for cancer patients based on radiomics and belief function theory," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 216–224, 2019.
4. J. M. Amigó, S. G. Balogh, and S. Hernández, "A Brief Review of Generalized Entropies," *Entropy*, vol. 20, 2018.
5. Y. Ma, Q. Liu, and Qian Z.-b., "Automated Image Segmentation Using Improved PCNN Model Based on Cross-entropy," Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.
6. S. Mannor, Peleg. D., R. Rubinstein, and et al., "The cross entropy method for classification," Proceedings of the 22nd international conference on Machine learning, 2005.

7. Z. Qu, J. Mei, L. Liu, and et al., "Crack Detection of Concrete Pavement With Cross-Entropy Loss Function and Improved VGG16 Network Model," 30th International Telecommunication Networks and Applications Conference (ITNAC), 2020.
8. L. Silva, J. Marques de Sá, and L. A. Alexandre, "Neural network classification using Shannon's entropy," ESANN 2005 Proceedings - 13th European Symposium on Artificial Neural Networks, 2005, pp. 217–222.
9. V. Rajinikanth, P. T. Krishnan, S. Satapathy, and et al., "Shannon's Entropy and Watershed Algorithm Based Technique to Inspect Ischemic Stroke Wound," Smart Intelligent Computing and Applications Proceedings of the Second International Conference on SCI 2018, November 2018, vol. 2.
10. U. Ruby and V. Yendapalli, "Binary cross entropy with deep learning technique for Image classification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, 10 2020.
11. D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, and J. Gonzalez-Rodriguez, "Deconstructing Cross-Entropy for Probabilistic Binary Classifiers," *Entropy*, vol. 20, no. 3, 2018.
12. M. Basseville, "Information: entropies, divergences et moyenne," Tech. Rep., INRIA, 2010.
13. Valeria Andreieva and Nadiia Shvai, "Generalization of cross-entropy loss function for image classification," *Mohyla Mathematical Journal*, vol. 3, pp. 3–10, 01 2021.
14. R Roselin et al., "Mammogram image classification: Non-shannon entropy based ant-miner," *International Journal of Computational Intelligence and Informatics*, vol. 4, 2014.
15. T. Brochet, J. Lapuyade-Lahorgue, S. Bougleux, M. Salaun, and S. Ruan, "Deep learning using havrda-charvat entropy for classification of pulmonary optical endomicroscopy," *IRBM*, vol. 42, no. 6, pp. 400–406, 2021.
16. S. Kumar and G. Ram, "A Generalization of the Havrda-Charvat and Tsallis Entropy and Its Axiomatic Characterization," *Abstract and Applied Analysis*, vol. 2014, pp. 1–8, 2014.
17. K Sirinukunwattana, E. Domingo, S. D. Richman, and et al., "Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning," *S: CORT consortium*, 2020.
18. J Noorbakhsh, S. Farahmand, P. Foroughi, and et al., "Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images.," *Nat Commun*, 2020.
19. S. Doppalapudi, R. G. Qiu, and Y. Badr, "Lung cancer survival period prediction and understanding: Deep learning approaches.," *Int J Med Inform.*, 2021.
20. X. Jiao, Y. Wang, F. Wang, and et al., "Recurrence pattern and its predictors for advanced gastric cancer after total gastrectomy," *Medicine (Baltimore)*, 2020.

# Convolutional Neural Network Classification of Liver Fibrosis Stages Using Ultrasonic Images Colorized by Features of Echo-Envelope Statistics



Akiho Isshiki, Dar-In Tai, Po-Hsiang Tsui , Kenji Yoshida,  
Tadashi Yamaguchi, and Shinnosuke Hirata 

**Abstract** The progression of liver fibrosis is the most important indicator that determines the prognosis of patients with diffuse liver disease. Variations in tissue structure triggered by liver fibrosis severely affect the texture and contrast of the ultrasound image. Therefore, progression can be non-invasively evaluated by analyzing ultrasound images. The convolutional neural network (CNN) classification of liver fibrosis stages using ultrasound images has also been studied. In previous studies, grayscale ultrasound images obtained using conventional ultrasound scanners were adopted as the input images. In this study, the modulation and colorization of the ultrasound images by the echo-envelope statistics that correspond to the texture and contrast of the ultrasound images have been proposed. In the proposed method, the colorized ultrasound image in RGB representation comprises the original image and two images modulated by different features of the echo-envelope statistics. Accordingly, the effect enhancement of tissue-structure variation by the colorization of the ultrasound images is promising in improving the accuracy of CNN classification. Therefore, CNN classification of the ultrasound images colorized by their 1st- and 3rd-order moments is demonstrated via the transfer learning of the VGG-16 pretrained network.

---

A. Isshiki

Department of Medical Engineering, Graduate School of Science and Engineering, Chiba University, 1-33 Yayoicho Inage-ku, Chiba Chiba 263-8522, Japan  
e-mail: [akiho.i@chiba-u.jp](mailto:akiho.i@chiba-u.jp)

D.-I. Tai

Department of Gastroenterology and Hepatology, Chang Gung Memorial Hospital, Linkou, No.5, Fuxing St, Guishan District, Taoyuan 33305, Taiwan

P.-H. Tsui

Department of Medical Imaging and Radiological Sciences, College of Medicine, Chang Gung University, No. 259, Wenhua 1st Road, Guishan District, Taoyuan 33302, Taiwan

K. Yoshida · T. Yamaguchi · S. Hirata (✉)

Center for Frontier Medical Engineering, Chiba University, 1-33 Yayoicho Inage-ku, Chiba Chiba 263-8522, Japan  
e-mail: [shin@chiba-u.jp](mailto:shin@chiba-u.jp)



**Keywords** Liver fibrosis · Quantitative diagnosis · Ultrasound image · Echo-envelope statistics · Convolutional neural network

## 1 Introduction

In the diffuse liver disease, the inflammation-necrosis-regeneration process of the liver parenchyma is repeated by chronic infections owing to HBV and/or HCV, alcoholic hepatitis and non-alcoholic steatohepatitis. This process is often associated with the irreversible fibrogenesis. Eventually, this disease leads to liver cirrhosis and hepatocellular carcinoma. The progression of liver fibrosis is the most important indicator that determines patient prognosis.

Liver fibrosis is quantitatively diagnosed via liver biopsy, ultrasound transient elastography (TE), and ultrasound shear-wave elastography (SWE). Pathological examination via liver biopsy remains the gold standard for determining the stage of liver fibrosis. However, liver biopsy is an invasive procedure often accompanied by complications [1]. In TE and SWE, shear waves are induced by a mechanical vibrator or acoustic radiation force impulse (ARFI) inside the liver. Liver elasticity is non-invasively estimated from the propagation speed of shear waves. However, inflammation or congestion other than fibrosis can also increase the liver elasticity [2].

The variation in tissue structure due to liver fibrosis severely affects the texture and contrast of the ultrasound image. Therefore, the progression of liver fibrosis can be non-invasively evaluated via ultrasound image analysis. Tissue characterization in the liver using echo-envelope statistics corresponding to the texture and contrast of the ultrasound images has been reported [3–6]. However, the classification of liver fibrosis stages by the convolutional-neural-network (CNN) analysis of ultrasound images has also been reported [7–9]. In previous studies on CNN classification, grayscale ultrasound images obtained using conventional ultrasound scanners were adopted as the input images. In this study, the modulation and colorization of the ultrasound images using echo-envelope statistics are proposed. The color image comprises three images in RGB representation. In the proposed method, the colorized ultrasound image comprises the original ultrasound images modulated by different features of the echo-envelope statistics. The effect enhancement of the tissue-structure variation in the ultrasound images by the colorization is promising in improving the accuracy of CNN classification. In this paper, the colorization of ultrasound images by moment, a feature of echo-envelope statistics, is described. Subsequently, the CNN classification of liver fibrosis stages using colorized ultrasound images is demonstrated via the transfer learning of the VGG-16, the pretrained CNN.

## 2 Method

### 2.1 Dataset

Clinical data of patients infected with HBV and/or HCV were obtained from the Chang Gung Memorial Hospital, Linkou, Taiwan. A portable ultrasound scanner and a convex array probe (Model 3000 and Model 5C2A, Terason, USA) were utilized to acquire the raw echo data. The center frequency of the transmitted ultrasound was 3.5 MHz, and the received echoes were stored by the sampling frequency of 30 MHz. The focal and maximum depths were fixed at 40 mm and 80 mm, respectively.

The stages of liver fibrosis were assessed by liver biopsy in accordance with the Metavir scoring system: normal liver (F0), early to severe fibrosis stages (F1–F3), and cirrhosis (F4). Because there were insufficient cases for the Metavir score of F0, the cases for the scores from F1 to F4 were used for CNN classification. The number of cases for each score was 20; in other words, 80 cases in total were used. The rate of intracellular fatty deposition, which was also assessed by liver biopsy, ranged from 0 to 30%.

### 2.2 Formation and Selection of Input Images

Ultrasound images were reconstructed by the scan conversion of the raw echo-data envelopes. The pixel spacings in lateral and depth directions of the image were 64.9  $\mu\text{m}$  and 63.3  $\mu\text{m}$ , respectively. In each image, the liver region was segmented by expert radiologists. Echo envelopes without logarithmic compression were normalized to eliminate the effects of focus and gain during transmission and reception. For the normalization of each pixel, the 2nd-order moment  $M_2$  of the echo envelopes in the region around the pixel was estimated as:

$$M_{2,i,j} = E[I_k^2], \quad (1)$$

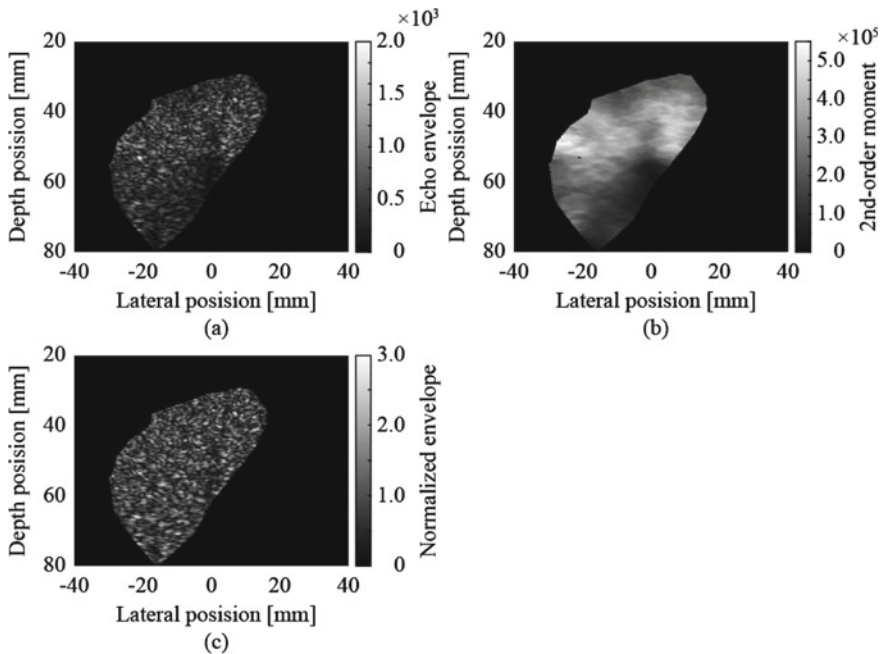
where  $i$  and  $j$  denote the coordinates of the center pixel of the region and  $I_k$  represents the  $k$ th pixel value in the region. The region was ellipse-shaped centering on the pixel and 12 ( $4 \times 3$ ) times the spatial resolution (1.9 mm  $\times$  2.4 mm) of the ultrasound image. In the normalization, each pixel value  $I_{ij}$  in the reconstructed ultrasound image was divided by the square root of the 2nd-order moment as follows:

$$\hat{I}_{ij} = \frac{I_{ij}}{\sqrt{M_{2,i,j}}}, \quad (2)$$

where  $\hat{I}$  denotes the normalized pixel value. Subsequently, the normalized pixels with values larger than 3 were discarded. The estimation-division-removal process

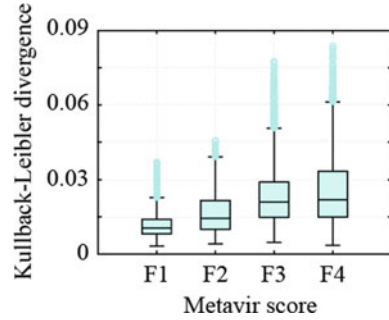
was repeated until the pixel with a value larger than 3 disappeared. Examples of the reconstructed ultrasound image, the distribution of the estimated 2nd-order moments, and the normalized ultrasound image are presented in Fig. 1. The regions of interest (ROIs), the input images to the CNN, were extracted from the normalized image within the ranges of  $-27$  to  $27$  mm and  $24$ – $64$  mm in lateral and depth sizes, respectively. The size of the ROI was  $15$  mm in lateral and depth ( $231 \times 237$  pixels). In the extracted ROI, pixels outside the liver region and the discarded pixels in the normalization were less than  $1\%$ . Sliding intervals were greater than  $1$  mm. All ROIs were rotated such that each vertical (depth) direction followed the scan line.

In total,  $27,324$  ROIs were extracted from  $80$  normalized ultrasound images. Several hundred ROIs were expected from each image (case). However, several echoes were assumed to be generated not from the liver parenchyma or fibrous tissues, but from vessel walls or lipid droplets in several ROIs. Therefore, the probability density functions (PDFs) of the ROIs were investigated. If most echoes are generated from homogeneous tissues, such as the liver parenchyma, the PDF of the ultrasound image can be approximated by a Rayleigh distribution. To compare the PDFs of the ROIs and the Rayleigh distribution, the Kullback–Leibler divergences (KLDs) between these distributions were calculated. A lower KLD value indicates that the PDF of the ROI is similar to the Rayleigh distribution. In other words, the echoes were mostly generated from the liver parenchyma. A higher KLD indicates that several



**Fig. 1** Reconstructed ultrasound image: **a**, the distribution of the estimated 2nd-order moments: **b**, the ultrasound image normalized by the square root of the 2nd-order moments: **c**

**Fig. 2** KLDs between the PDFs of the extracted ROIs (echo data) and Rayleigh distribution



echoes generated from the fibrous tissues are mixed. The KLDs between the PDFs of all ROIs and the Rayleigh distribution are presented in Fig. 2. The high-KLD ROIs of F1 and low-KLD ROIs of F4, respectively, appear and should be discarded because the texture and contrast of the ROIs may not correspond to the tissue structures.

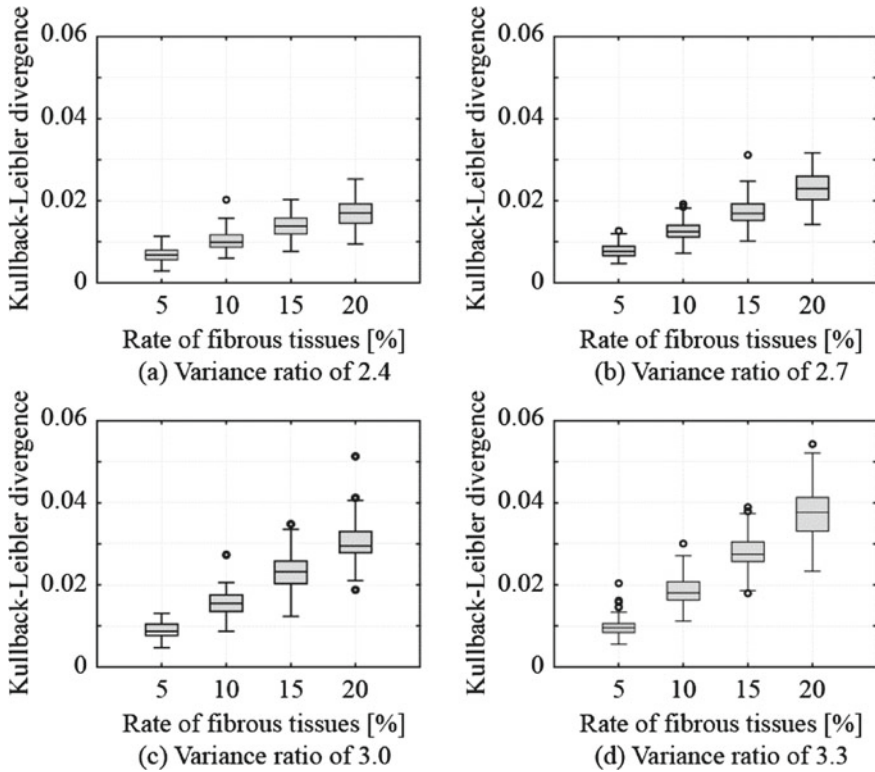
To calculate the KLD between the PDF corresponding to the assumed tissue structure and the Rayleigh distribution, the multi-Rayleigh model was introduced [5]. In the model, the PDF of the echoes from the fibrotic liver is expressed by the combination of the low-variance Rayleigh distribution corresponding to the echoes from the liver parenchyma and high-variance Rayleigh distribution corresponding to those from fibrous tissues. In the calculation of KLDs based on the multi-Rayleigh model, the rates of fibrous tissues were set to 5, 10, 15, and 20% in the cases of F1, F2, F3, and F4, respectively. The variance ratios between both Rayleigh distributions were set from 2.4 to 3.3 in all cases. The KLDs calculated from the random numbers of 5000 samples in each PDF are presented in Fig. 3. In each score in Fig. 2, the ROIs whose KLDs were from the lower whisker of the variance ratio of 2.4 to the higher whisker of the variance ratio of 3.3 were selected. Furthermore, 2560 ROIs in each score, approximately 128 ROIs in each case, were adopted for the CNN classification. The KLDs between the PDFs of the used ROIs and the Rayleigh distribution are presented in Fig. 4.

### 2.3 Modulation and Colorization of Input Images

The 1st- and 3rd-order moments were employed as the features of the echo-envelope statistics for the colorization of the ROIs. In each ROI, the 1st-order moment  $M_1$  and 3rd-order moment  $M_3$  of the normalized pixel values were estimated as follows:

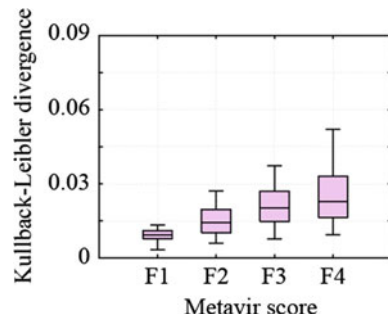
$$M_1 = E[\hat{I}_n], \quad (3)$$

$$M_3 = E[\hat{I}_n^3], \quad (4)$$



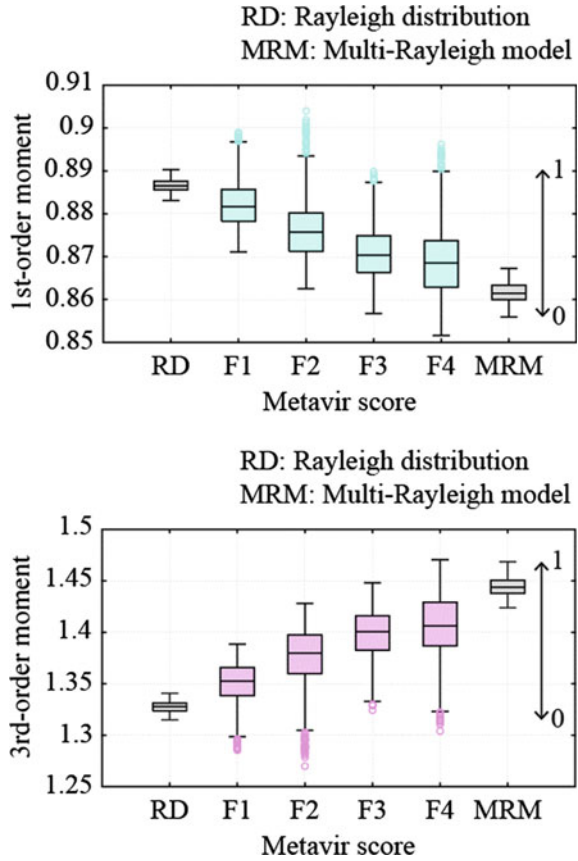
**Fig. 3** KLDs between the PDFs of the multi-Rayleigh models (random number) and the Rayleigh distributions

**Fig. 4** KLDs between the PDFs of the used ROI (echo data) for the CNN classification and the Rayleigh distribution



where  $\hat{I}_n$  denotes the  $n$ th normalized pixel value in the ROI. The estimated moments of all ROIs utilized for the CNN classification are presented in Fig. 5. Before the modulation, the moments were normalized from 0 to 1 based on the calculated moments from the random numbers of 5000 samples. Regarding the 1st-order moments, the normalization band was set from the lower whisker of the variance ratio of 3.3 and

**Fig. 5** The 1st- and 3rd-order moments of the utilized ROIs (echo data) and their normalized bands between the moments estimated by the multi-Rayleigh model and the Rayleigh distribution

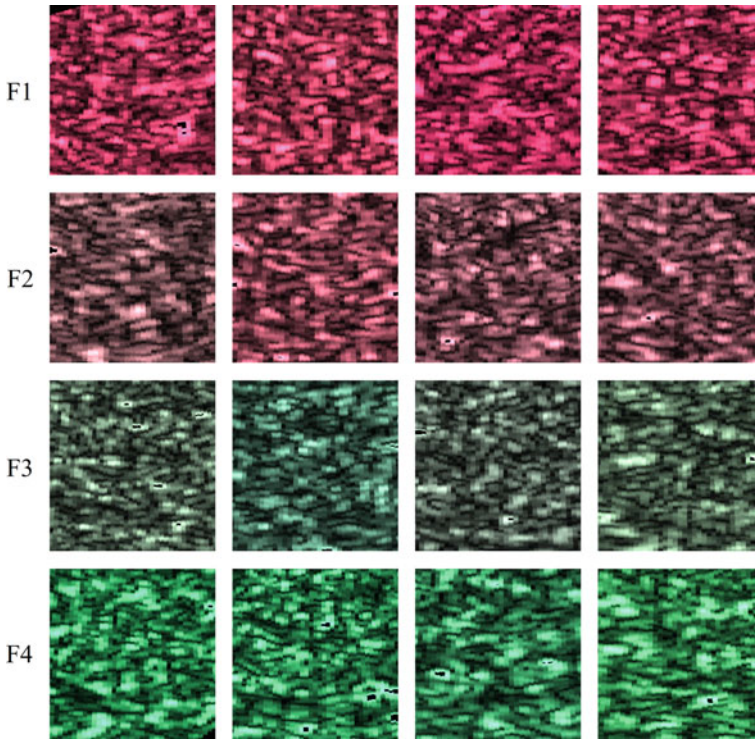


fibrous-tissue rate of 20% to the higher whisker of the variance ratio of 1 (Rayleigh distribution). Regarding the 3rd-order moments, which were set from the lower whisker of the Rayleigh distribution to the higher whisker of the variance ratio of 3.3 and fibrous-tissue rate of 20%. The normalized pixel values in the ROI were modulated as:

$$I_{1,i,j} = \hat{I}_{i,j} \times 2^{2 * (\widehat{M}_1 - 0.5)}, \tag{5}$$

$$I_{3,i,j} = \hat{I}_{i,j} \times 2^{2 * (\widehat{M}_3 - 0.5)}, \tag{6}$$

where  $\widehat{M}_1$  and  $\widehat{M}_3$  represent the normalized moments from 0 to 1. The colored image input to the CNN was created using the original image of  $\hat{I}_{i,j}$ , modulated image  $I_{1,i,j}$ , and that of  $I_{3,i,j}$  in the blue, red, and green layers, respectively. Examples of the colored ROIs are presented in Fig. 6.



**Fig. 6** Colorized ROIs comprising the original images, the images modulated by the 1st-order moments and images modulated by the 3rd-order moments in the blue, red, and green layers, respectively

## 2.4 Learning and Validation of Networks

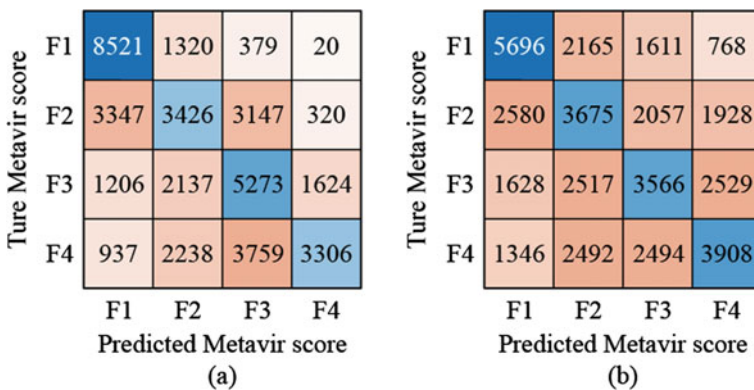
The selected ROIs and the horizontally flipped ROIs (20,480 ROIs in total) were used to learn and validate the network. In this study, the pretrained network VGG-16 in the Deep Learning Toolbox in MATLAB is utilized for the CNN classification of the liver fibrosis stages. The VGG-16 comprises 13 convolution layers and 3 fully connected layers. For the classification of the Metavir scores F1–F4, the last fully connected layer was replaced with the new fully connected layer (input: 4096, output: 4). The weights of the layer were initialized using random numbers. In the transfer learning, only the last two convolutional layers and three fully connected layers were trained to prevent early overfitting. The training was performed using the stochastic gradient descent with the mini-batch processing of 64 images. The dropout between the fully connected layers were 80%, the learning rate was  $3.5 \times 10^{-5}$ , and the number of epochs were 4 and 5 in the cases of the colorized ROIs and grayscale ROIs, respectively. Averages of validation losses were minimized at the epochs. To validate the trained network, five-fold cross-validation was performed. All ROIs

were divided into five sets. Three of sets were adopted to train the network and the remaining two sets were adopted to validate the network. The combination of these sets was switched and repeated five times. Therefore, each ROI was adopted for learning thrice and for validation twice.

### 3 Result and Discussion

The ROIs colored by the 1st- and 3rd-order moments as the features of the echo-envelope statistics were classified by the VGG-16 trained by the ROIs. For comparison, the original grayscale ROIs indicated that the normalized echo envelopes were also classified by the VGG-16 trained using the original ROIs. The results of the confusion matrices for both cases are presented in Fig. 7. Regarding the colored ROIs, the accuracies of the predicted Metavir scores in F1, F2, F3, F4, and total were 83.2%, 33.5%, 51.5%, 32.3%, and 50.1%, respectively. Regarding the original ROIs, the accuracies were 55.6%, 35.9%, 34.8%, 38.2%, and 41.1%, respectively. The accuracy of the classification was improved 9% via the colorization of the ROIs. Limited to the classification of F1–F2 and F3–F4, the accuracy was improved from 65 to 75%. Furthermore, the ROIs of F1 predicted as F3 or F4, and those of F2 predicted as F4, were significantly decreased. However, in the ROIs of F4, the accuracy of the classification was not significantly improved; in other words, the ROIs predicted as F1 and F2 were not significantly decreased.

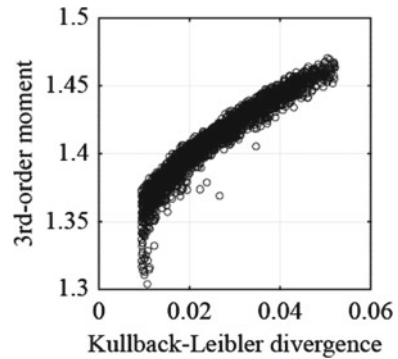
In Fig. 5, 3rd-order moments of the several ROIs of F4 are lower than those of F1 or the Rayleigh distribution despite their KLD with the Rayleigh distribution are higher than those of F1. The 3rd-order moments and KLDs of the F4 ROIs are presented in Fig. 8. Basically, as the 3rd-order moment increases, KLD also increases in the multi-Rayleigh model. When the 3rd-order moments go below that of the Rayleigh



**Fig. 7** Confusion matrices of the CNN classification of the liver fibrosis stages using the colored ROIs: **a**, original grayscale ROIs; **b**



**Fig. 8** The 3rd-order moments and KLDs with the Rayleigh distribution of the utilized ROIs of F4



distribution, the KLD increases from 0.0094, the bottom range of the KLD in F4. The tissue structures in these ROIs may not correspond to the multi-Rayleigh model. Therefore, such ROIs in all Metavir scores should be preliminarily discarded.

## 4 Conclusion

For the non-invasive and accurate quantitative diagnosis of liver fibrosis caused by the diffuse liver disease, the CNN analysis of ultrasound images colorized by the features of echo-envelope statistics was proposed. In this study, the CNN classification of liver fibrosis stages using ultrasound images colorized by the 1st- and 3rd-order moments of echo envelopes was demonstrated for clinical ultrasound images of patients infected with HBV and/or HCV. The accuracy of the predicted Metavir scores could be improved from 41.1 to 50.1% by the colorization. In this study, the ROIs utilized for the CNN classification were selected using their KLDs with the Rayleigh distribution. However, moments of several ROIs were not appropriate, although KLDs were appropriate for their liver fibrosis stages. Therefore, discarding such ROIs may improve the accuracy.

**Acknowledgements** This study was partly supported by the Takahashi Industrial and Economic Research Foundation.

## References

1. Seeff, L. B., Everson, G. T., Morgan, T. R., *et al.*: Complication rate of percutaneous liver biopsies among persons with advanced chronic liver disease in the HALT-C Trial. *Clinical Gastroenterology and Hepatology* 8(10), 877–883 (2010).

2. Millong, G., Friedrich, S., Adolf, S., *et al.*: Liver stiffness is directly influenced by central venous pressure. *Journal of Hepatology* 52(2), 206–210 (2010).
3. Weng, L., Reid, J. M., Shankar, P. M., *et al.*: Ultrasound speckle analysis based on the K distribution. *The Journal of the Acoustical Society of America* 89, 2992–2995 (1991).
4. Dutt, V., Greenleaf, J. F.: Ultrasound echo envelope analysis using a homodyned K distribution signal model. *Ultrasonic Imaging* 16(4), 265–287 (1994).
5. Mori, S., Hirata, S., Yamaguchi, T., *et al.*: Quantitative evaluation method for liver fibrosis based on multi-Rayleigh model with estimation of number of tissue components in ultrasound B-mode image. *Japanese Journal of Applied Physics* 57(7S1) 07LF17 (2018).
6. Fang, F., Li, Q., Tai, D.-I., *et al.*: Ultrasound assessment of hepatic steatosis by using the double Nakagami distribution: A feasibility study. *Diagnostics* 10(8), 557 (2020).
7. Meng, D., Zhang, L., Cao, G., *et al.*: Liver fibrosis classification based on transfer learning and FCNet for ultrasound images. *IEEE Access* 5, 5804–5810 (2017).
8. Lee, J. H., Joo, I., Kang, J. W., *et al.*: Deep learning with ultrasonography: automated classification of liver fibrosis using a deep convolutional neural network. *European Radiology* 30, 1264–1273 (2019).
9. Saito, R., Koizumi, N., Nishiyama, Y., *et al.*: Evaluation of ultrasonic fibrosis diagnostic system using convolutional network for ordinal regression. *International Journal of Computer Assisted Radiology and Surgery* 16, 1969–1975 (2021).

# FedRNN: Federated Learning with RNN-Based Aggregation on Pancreas Segmentation



Zengtian Deng, Touseef Ahmad Qureshi, Sehrish Javed, Lixia Wang, Anthony G. Christodoulou, Yibin Xie, Srinavas Gaddam, Stephen Jacob Pandol, and Debiao Li

**Abstract** Federated learning (FL) has been applied by several studies for pancreas segmentation. However, handling heterogeneous datasets across participating sites remains to be a challenge. To address the heterogeneity issues to further improve the performance of FL, we developed an innovative aggregation method, FedRNN, which used a Recurrent Neural Network (RNN) to adjust the aggregation weight of each site's model based on the history of model loss and aggregation weight. At each round, the RNN took in the previous round aggregation weight and current round loss value from each site to estimate the optimal aggregation weight for the current round. Additionally, Mean Square Error (MSE) was applied for balanced performance across the clients. Based on cross-site validation, FedRNN outperformed the existing FL algorithms with an overall mean dice score of 78.7% and was up to 4.2% in improvement. In addition, FedRNN had the most stable performance across all clients in terms of the lowest standard deviation. Based on the results, the loss and aggregation weight history can be beneficial to the aggregation process of FL. Additionally, since FedRNN does not have restrictions on the form of loss functions, it can be applied to other tasks such as classification and object detection.

**Keywords** Federated learning · Model aggregation · Pancreas segmentation · Recurrent neural network

## 1 Introduction

To date, Deep Learning (DL) has been widely implemented in the medical field for various tasks. Although it is very promising in some circumstances, adequate training data is not always available at a single institution, necessitating collaboration

---

Z. Deng (✉) · T. A. Qureshi · S. Javed · L. Wang · A. G. Christodoulou · Y. Xie · S. Gaddam · S. J. Pandol · D. Li  
Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA  
e-mail: [Zengtian.Deng@cshs.org](mailto:Zengtian.Deng@cshs.org)

Z. Deng · A. G. Christodoulou · S. Gaddam · S. J. Pandol · D. Li  
University of California Los Angeles, Los Angeles, CA 90095, USA

between multiple. However, data sharing across institutions faces technical difficulties transferring the data and has privacy concerns on disclosing patient information [1]. Therefore, an efficient training method without data sharing while preserving privacy is highly desirable.

In recent years, federated learning (FL), an AI framework that allows private and decentralized model training, has been applied to multiple medical research and projects. In the field of medical segmentation, FL has been applied for multi-site brain segmentation [2], brain tumor segmentation [3, 4], and pancreas segmentation [5–7], etc. In general, FL has the same model separately trained at each client site, and the updates of trained models reflecting the dataset of each client are aggregated to generate the new model. One of the commonly used federated learning schemes is centralized federated learning [1], which has a server to distribute the model to each client and aggregate model updates.

Model aggregation weight, the factor that scales the model update of each client before the aggregation process, is crucial to the end form of the model. In the first FL paper, McMahan et al. proposed Federated Averaging (FedAvg) as the first aggregation algorithm for FL [8]. FedAvg applies fixed aggregation weights based on the size of client's training data, which is lightweight and easy to implement. However, FedAvg cannot handle heterogeneous datasets well. To incorporate data heterogeneity into consideration, Shen et al. evaluated various algorithms and discovered that Dynamic Weight Averaging (DWA) yielded the highest performance [6]. DWA adjusts aggregation weight based on the improvement of training loss between consecutive rounds and has been shown to have better performance than FedAvg. However, DWA is not robust to fluctuations in the loss over the training history and does not take the aggregation weight setting into consideration. Therefore, to fully utilize the temporal information during the training process, we designed an innovative aggregation algorithm called FedRNN that uses a recurrent neural network (RNN) to incorporate both the loss and the aggregation weights into the aggregation process.

Pancreatic ductal adenocarcinoma (PDAC), which constitutes over 90% of pancreatic cancer, is one of the highly lethal malignancies [9], and pancreas imaging is an initial and key study for diagnosis and management. Thus, pancreas segmentation is the critical step in the process of improving imaging methods in the field. However, since manual segmentation could be timing consuming, automatic segmentation is preferable and an ongoing research field for both centralized learning [10, 11] and federated learning [6, 7]. Therefore, we chose multi-site pancreas segmentation as the task, aiming to prove the feasibility of FedRNN and further improve existing FL performance on pancreas segmentation.

An important consideration in FL is its performance compared to centralized learning. According to Nilsson et al., with heterogeneous setting of the MNIST dataset, FedAvg performs slightly worse than all-dataset centralized learning [12]. However, no previous FL studies on pancreas segmentation have compared all-dataset centralized learning with FL methods. Thus, we also performed all-dataset centralized learning for comparison with FL methods.

Our research work was summarized into the following contributions:

- (1) We introduced a new FL algorithm, FedRNN, which used RNN to adjust the aggregation weight based on the loss and aggregation changed over time.
- (2) We visualized and explored the relationship between aggregation weight and local training loss for both DWA and FedRNN methods.
- (3) We compared the performance of FL algorithms with both the performance of CL algorithms using single datasets and the CL algorithm using all datasets.

## 2 Materials and Methods

### 2.1 Dataset

The study includes three publicly available datasets. The first dataset is from Task 07 of Medical Segmentation Decathlon (MSD)<sup>1</sup> [13] and has 281 CT abdominal scans; The second dataset is the Pancreas-CT dataset with 82 scans published by the National Institute of Health Clinical Center (NIH)<sup>2</sup> [14]; The third dataset is from MICCAI's Multi-Atlas Labeling beyond the Cranial Vault challenge (BTCV)<sup>3</sup> [15] and has 30 abdominal scans. All three datasets are publicly available. Each case is resampled to the dimension of  $1.0 \text{ mm}^3 \times 1.0 \text{ mm}^3 \times 2.0 \text{ mm}^3$  using bilinear interpolation. The density range of each scan was cropped within the range from  $-200$  to  $250$  in Hounsfield Unit and normalized. Random density shift and random affine transform were applied for data augmentation. Additionally, RandCropByPosNegLabel from MONAI [16] was used to randomly crop each case volume to multiple volumes of size  $112 \times 112 \times 48$ . For our study, 4 volumes were generated per case, and the center voxels of the cropped volumes were enforced to be evenly distributed to be arbitrary positive and negative labels.

### 2.2 Federated Learning Framework

As shown in the centralized FL setup in Fig. 1, at the start of each round, the server first distributes the global model to each client. The clients then evaluate the received global model for cross-site validation, train the model with local datasets, and upload the trained model updates processed by the privacy protocol back to the server for aggregation. The aggregated model then becomes the new global model for the next round. The goal of FL can be expressed as minimizing the function shown in Eq. 1, where the global loss function  $L$  is comprised of the weighted sum of the individual loss function  $L_k$  at each client  $k$  [1];  $\theta$  is the model parameter; and  $X_k$  and  $\omega_k$  are the input data and the aggregation weight of client  $k$ , respectively.

---

<sup>1</sup> <http://medicaldecathlon.com>.

<sup>2</sup> <https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT>.

<sup>3</sup> <https://www.synapse.org/#!/Synapse:syn3193805/wiki/89480>.

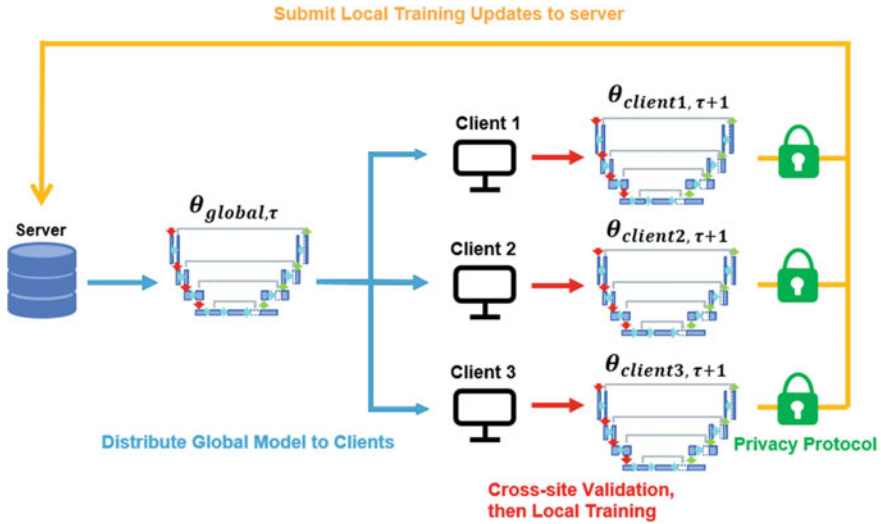


Fig. 1 Overview of centralized FL framework

$$\min_{\theta} L(X, \theta) = \min_{\theta} \sum_{k=1}^K \omega_k L_k(X_k, \theta) \tag{1}$$

### 2.3 Model Aggregation

Prior studies have used various approaches for model aggregation. For the FedAvg [8] method,  $\omega_k$  is selected to be the ratio of local data size to the total data size as shown in Eq. 2, where  $n_k$  is the amount of data at client  $k$  and  $n_{total}$  is the total amount of data across all datasets.

$$\omega_k = \frac{n_k}{n_{total}} \tag{2}$$

To account for data heterogeneity, DWA [6] adaptively adjusted the aggregation weight to focus on the model that suppresses more training loss on local dataset, as expressed in Eq. 3, where  $\xi$  and  $T$  are the scaling factors, and  $\rho_{k, \tau-1}$  is the loss difference equation of client  $k$  at time  $\tau - 1$  defined in Eq. 4.

$$\omega_{k, \tau} = \frac{\xi \exp(\rho_{k, \tau-1}/T)}{\sum_{i=1}^K \exp(\rho_{i, \tau-1}/T)} \tag{3}$$

$$\rho_{k,\tau-1} = \frac{\mathcal{L}_{k,\tau-1}}{\mathcal{L}_{k,\tau-2}} \tag{4}$$

To broaden the contribution of the loss factor and incorporate aggregation weight into consideration, we proposed FedRNN, which uses the recurrent neural network (RNN) to adaptively estimate the appropriate aggregation weight. As shown in Eq. 5,  $\omega_\tau$  is the vector of aggregation weight of all clients at time  $\tau$ , and  $\mathcal{L}_{seg,\tau}$  is the vector of training loss from all clients at time  $\tau$ .

$$\omega_\tau = f_{RNN}(\omega_{\tau-1}, \mathcal{L}_{seg,\tau}) \tag{5}$$

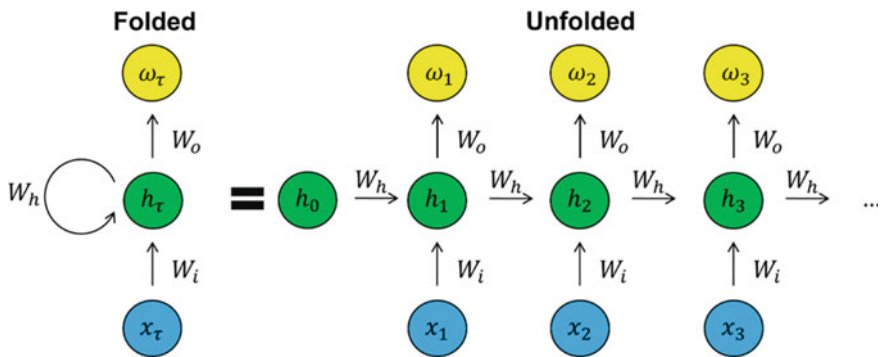
The RNN model used in this study is Elman RNN [17], and Eq. 5 can be extended to Eqs. 6–8, where  $x_\tau$  is the input of the model at time  $\tau$ , and  $h_\tau$  is the hidden state feature at time  $\tau$ . As shown in Eq. 8,  $x_\tau$  is the previous-round aggregation weight and the current round segmentation loss connected through additive attention. The SoftMax function is applied to the model output to enforce the aggregation weights to sum to one as shown in Eq. 6. The overall RNN structure is depicted in Fig. 2.

$$\omega_\tau = \sigma_{SoftMax}(W_o h_\tau + b_o) \tag{6}$$

$$h_\tau = \sigma_{tanh}(W_i x_\tau + b_i + W_h h_{\tau-1} + b_h) \tag{7}$$

$$x_\tau = \omega_{\tau-1} + \mathcal{L}_{seg,\tau} \tag{8}$$

The optimization goal of the RNN model is to minimize the segmentation loss at each client. However, the backpropagation from local training loss to the aggregation



**Fig. 2** The structure of RNN network and its equivalent unfolded form that shows the update progress of RNN along the input sequence.  $\tau$  represents the time point or round number;  $x_\tau$  represents the input of the model;  $h_\tau$  represents the hidden state of the RNN network; and  $\omega_\tau$  represents the output of the model, which is the aggregation weights to be used for next round

weight will be computationally expensive and prone to vanishing gradient. Therefore, we designed an innovative loss function to approximate the overall process as depicted in Eq. 9, where  $\theta_{RNN}$  is the model parameter of RNN network;  $\omega'_\tau$  is the optimal aggregation weight, and  $\mathcal{L}'_{seg,\tau}$  is the segmentation loss.

$$\min_{\theta_{RNN}} \mathcal{L}_{RNN} = \min_{\theta_{RNN}} \left\| \omega'_\tau \mathcal{L}'_{seg,\tau} - f_{RNN}(\omega_{\tau-1}, \mathcal{L}_{seg,\tau}, \theta_{RNN}) * \mathcal{L}_{seg,\tau} \right\|^2 \quad (9)$$

In Eq. 9,  $f_{RNN}(\omega_{\tau-1}, \mathcal{L}_{seg,\tau}, \theta_{RNN}) * \mathcal{L}_{seg,\tau}$  can be interpreted as the skip-connection-like multiplication to the final optimization loss of the RNN model. Additionally, since the optimal segmentation loss is zero, Eq. 9 can be rewritten into the form in Eq. 10, and the optimization process will be minimizing the product between RNN output and the induced segmentation loss. Mean Squared Error is used for the RNN model to enforce balanced performance across the clients.

$$\min_{\theta_{RNN}} \mathcal{L}_{RNN} = \min_{\theta_{RNN}} \left\| f_{RNN}(\omega_{\tau-1}, \mathcal{L}_{seg,\tau}, \theta_{RNN}) * \mathcal{L}_{seg,\tau} \right\|^2 \quad (10)$$

Theoretically, the hidden states of the RNN model store information about the history of aggregation weight and segmentation loss. Assuming there is a latent space of model parameters, and heterogeneous datasets have varied optimal model state, the losses from each site can be interpreted as the distance to the optimal model of each local dataset; the aggregation weights represent the step size of the model during optimization; and the model updates imply the direction towards the specific optimal model. Then, the overall optimization goal of RNN can be interpreted as finding the optimal position in the model space that has the shortest distance to all optimal models.

## 2.4 Experimental Setup

The experiment was run using *NVIDIA container for Pytorch*, release 21.02,<sup>4</sup> with Clara Train 4.0<sup>5</sup> [18] as the package for federated learning framework. The percentile privacy protocol used in this study only allows 25% of the largest model updates in absolute value to be sent from clients to the server. The training process was performed within a single server using different containers, and each client was assigned a GeForce RTX 2080 Ti with around 4800 MB memory used per client.

The 3D UNet with residual unit is used as the base model in this study [19]. The 3D UNet is a 4-layer network with the scaling factor as two. Two residual units are used per layer. The model is implemented with project MONAI package [16]. To balance model performance and communication efficiency, the training was set to have two local epochs and 500 aggregation rounds. All CL methods were trained

<sup>4</sup> [https://docs.nvidia.com/deeplearning/frameworks/pytorch-release-notes/rel\\_21-02.html](https://docs.nvidia.com/deeplearning/frameworks/pytorch-release-notes/rel_21-02.html).

<sup>5</sup> <https://docs.nvidia.com/clara/clara-train-sdk/index.html>.



for 1000 epochs with the exact same setting. The models were trained using Adam optimizer with the initial learning rate at  $2e^{-4}$  and subsequent learning rate adjusted by  $lr * (1 - n_{epoch}/n_{max\_epoch})^{0.9}$ . A combination of Dice Loss and Focal Loss is implemented on a scale of 1:1 to cope with class imbalance. Batch normalization is used, and the model is trained with 0.5 drop off rate. The RNN model is jointly trained with the segmentation model using the training loss, and the validation dataset is used to prevent overfitting of the whole framework combining RNN and 3D UNet. The output of the model is the vector of probabilities for each voxel being pancreas and background. During inference, sliding window inference from MONAI is used to predict the pancreas volume of each case, and the SoftMax function is applied to the output to ensure the probabilities of each voxel sum up to one. Then, both the ground truth and the model output are converted to one-hot encoding for comparison. In case both probabilities are the same, the voxel will be considered as the background.

As the ablation study on the amount of effective temporal information for RNN, the input sequence to RNN was experimented with 10, 15 and 20. Additionally, to study the influence from RNN complexity, the size of hidden feature was tested with 5, 7, and 10. To ensure enough input for RNN, the aggregation weights were fixed to same value as FedAvg until enough aggregation weights and losses are collected.

### 3 Results

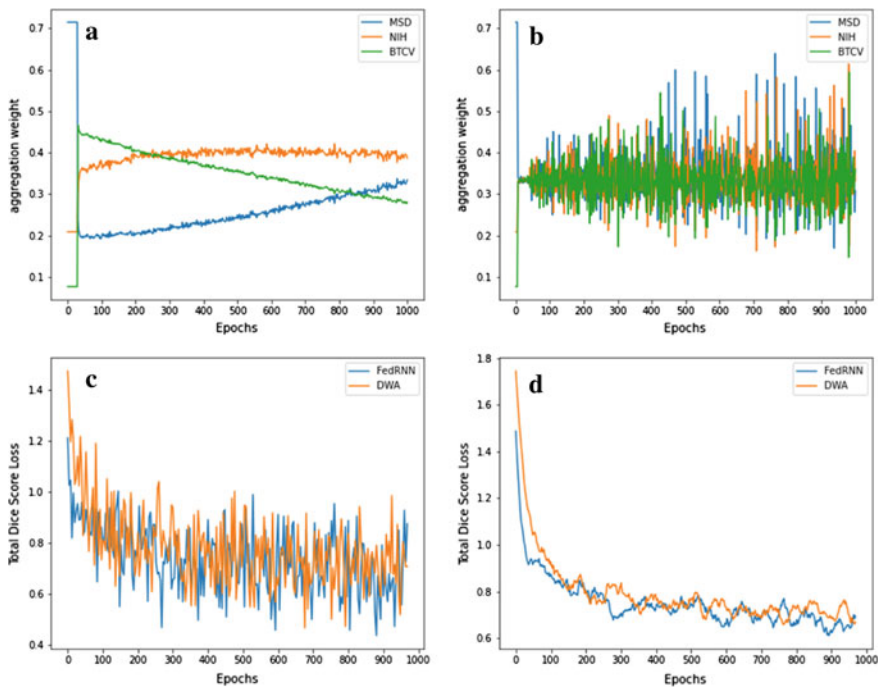
Cross-site validation in terms of Dice Similarity Coefficient (DSC) was performed on both CL methods and FL methods as shown in Table 1. Besides the performance on each site, the mean DSC of all test cases (Case Avg) was also evaluated. Paired Samples T-Test was performed between datasets. Based on the cross-site validation, CL with MSD dataset outperformed the other two CL methods. However, CL with NIH dataset performed the best on its own test set. From the local CL result, a positive

**Table 1** Cross-site validation on average dice similarity coefficient (DSC) of all methods with standard deviation. The best DSC of FL methods is shown in bold

Method	MSD (93)	NIH (27)	BTCV (10)	Case avg
<i>Centralized learning</i>				
MSD local	74.8 ± 13	69.8 ± 15	77.4 ± 6.2	73.9 ± 13
NIH local	60.0 ± 22	73.3 ± 12	69.2 ± 12	63.4 ± 21
BTCV local	44.1 ± 22	56.8 ± 19	56.9 ± 23	47.7 ± 22
All dataset local	76.5 ± 12	79.4 ± 5.6	80.7 ± 4.8	77.4 ± 10
<i>Federated learning</i>				
FedAvg	75.2 ± 11	72.2 ± 7.6	74.6 ± 7.0	74.5 ± 10
DWA	77.5 ± 11	77.6 ± 6.7	80.1 ± 5.0	77.7 ± 11
FedRNN	<b>78.1 ± 11</b>	<b>80.1 ± 5.5</b>	<b>80.8 ± 5.7</b>	<b>78.7 ± 9.7</b>

correlation between the model performance and the size of the dataset was seen, and CL with all datasets yielded the best performance among all CL methods. As shown in Table 1, all FL methods performed better than local CL methods. Among all FL methods, FedRNN outperformed both the DWA by 1.0% ( $P < 0.001$ ) and FedAvg by 4.2% ( $P < 0.001$ ) in terms of Case Avg. Comparing to CL with all data, FedRNN performed better than CL with all data ( $P < 0.001$ ); DWA had similar performance with CL with all data ( $P = 41.1$ ); and FedAvg had worse performance than CL with all data ( $P < 0.001$ ). In addition, FedRNN also had the lowest standard deviation on Case Avg compared to other methods.

Figure 3 showed the aggregation weight and loss change of FedRNN and DWA during the training process. As shown in Fig. 3a, FedRNN had the aggregation weights slowly adjusted to similar values during the training process. On the contrary, as shown in Fig. 3b, the aggregation weights of DWA showed arbitrary fluctuations and increasing difference between each other, which conformed with the plot shown in the work of Shen et al. [6]. Figure 3c showed the loss plots of both methods, and the two plots were indistinguishable from each other. However, since the best performed FedRNN model used the previous 15 rounds for each prediction, the 15-round moving average plot was created shown in Fig. 3d. As shown in Fig. 3d, the loss of FedRNN was explicitly lower than that of DWA for most of the time.



**Fig. 3** The aggregation weight plot of best performed FedRNN **a** and DWA **b** as well as their loss plot **c** and the loss plot in 15-round moving average **d** during the training process

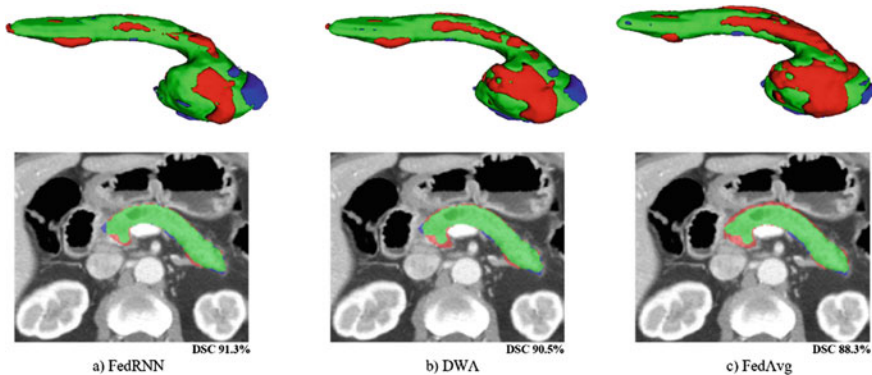
Additionally, the loss of FedRNN also dropped faster than that of DWA during the first 100 epochs.

Table 2 showed the cross-site validation on the ablation study of FedRNN. As shown in Table 2, the best performed FedRNN model had a sequence length of 15 and 10 hidden features. For the RNN models with better performance, a certain ratio between sequence length and hidden feature size was maintained, such as group 2 and group 4. The performance immediately dropped when either parameter was set above the optimal setting, as shown in group 3 and group 5.

Figure 4 showed the example pancreas segmentation of the three FL methods from the MSD test set in both 3D volume and 2D slice view. The green volume represented the overlap between ground truth label and prediction (True Positive); the red volume represented the error prediction (False Positive); and the blue volume represented the unidentified ground truth label (False Negative). Compared to DWA and FedAvg, FedRNN had explicitly less false positive voxels and better DSC.

**Table 2** Cross-site validation on ablation study of FedRNN in terms of hidden feature and input sequence length. Best performance is shown in bold

Group	Sequence length	Hidden feature	MSD (93)	NIH (27)	BTCV (10)	Case avg
1	10	5	78.1 ± 11	78.9 ± 6.8	80.7 ± 5.7	78.1 ± 9.3
2	10	7	<b>78.2 ± 11</b>	78.8 ± 6.7	80.1 ± 6.2	78.5 ± 10
3	10	10	76.7 ± 12	77.0 ± 9.6	79.5 ± 6.2	77.0 ± 11
4	15	10	78.1 ± 11	<b>80.1 ± 5.5</b>	<b>80.8 ± 5.7</b>	<b>78.7 ± 9.7</b>
5	20	10	77.8 ± 12	78.2 ± 8.5	80.0 ± 6.3	78.0 ± 11



**Fig. 4** Example pancreas segmentation by FedRNN (Left), DWA (Middle), and FedAvg (Right). The upper row displays the 3D volume of the segmentation, and the lower row shows the example 2D axial slices of the segmentation with dice score shown. For all plots, the green volume represents the overlap between ground truth and prediction (True Positive); the blue volume represents the ground truth volume not detected by the model (False Negative); and the red volume represents the error prediction volume (False Positive)

## 4 Discussion

Based on the cross-site validation result, FedAvg showed better performance on MSD dataset because it has the largest number of training data. In addition, comparing CL with all datasets to FL methods, all-dataset CL performed better than FedAvg, which conforms with the observation from Nilsson et al. [9]. However, DWA performed similar to the all-dataset CL method, and FedRNN performed better than all-dataset CL method. This might be due to the difference between the training scheme of CL and FL. For all-dataset CL, the model is trained sequentially with data from all datasets; while for the FL setting, the model is parallelly trained on the data from different groups. The results may imply that for heterogeneous data, parallel training of the model on different groups could yield better results than traditional sequential training.

Based on Table 2, the input sequence length and the hidden feature jointly influenced the result of FedRNN, and a correct ratio between the two is crucial based on the results. On the other side, RNN with inadequate combination may either not have enough information or not have enough model capacity. Additionally, larger hidden feature with longer sequence length yields better result, as shown in group 2 and 4 from Table 2.

Based on Fig. 3b, the aggregation weight of DWA had an unstable trend due to its dependence on the fluctuation of the loss. On the contrary, FedRNN had minimum interference from the fluctuation of loss by taking the history of loss and aggregation weight into account. Same with DWA, FedRNN will focus on the aggregation weights of the model that has less loss. However, the MSE loss and the history of loss and aggregation weight help stabilize the distribution of aggregation weight to force the model paying attention to updates of the model with larger loss, yielding a better result than DWA.

The performance of FedRNN also shows better performance visually than other FL methods. As shown in Fig. 4, both DWA and FedAvg tend to overpredict the pancreas region, resulting in a much larger false positive red volume comparing to FedRNN. However, all three FL methods fail to detect part of the pancreas head, leaving a blue False Negative volume in Fig. 4. Since all FL frameworks use the Residual UNet with the exact same setting, such errors could be due to the insufficient capability of the segmentation model itself. Therefore, experimenting with more complicated model structure and developing data harmonization methods could be possible directions for future research.

Since the RNN network only takes the aggregation weights and the averaged loss from each site as input, FedRNN is still simple and lightweight. In terms of communication, FedRNN only requires the average loss value of each site to be transferred to the server. In addition, since FedRNN does not have restrictions on the type of loss to be learned except for a fixed ideal value of the loss function output, which is usually zero. Therefore, FedRNN can smoothly fit into other AI-related tasks like image classification, registration, super-resolution etc.

## 5 Conclusion

We developed a new federated learning algorithm, FedRNN, which uses an RNN to extract the temporal information of aggregation weights and losses to automatically adjust the model aggregation in FL for better model performance. It not only showed better DSC than FedAvg, DWA and even all-dataset CL, but also had less variation in terms of DSC standard deviation across data from all sites. Additionally, since FedRNN has no restriction on the loss function and model type, it can be easily applied to other deep learning tasks. More advanced data harmonization methods and other more complicated recurrent networks such as Long Short-Term Memory (LSTM) networks may further improve the generalization and accuracy of the algorithm.

## References

1. Rieke, N., Hancox, J., Li, W. et al.: The future of digital health with federated learning. *npj Digit. Med.* 3, 119 (2020).
2. Roy, A. G., Siddiqui, S., Pölsterl, S., Navab, N. & Wachinger, C. Braintorrent: a peer-to-peer environment for decentralized federated learning. arXiv preprint [arXiv:1905.06731](https://arxiv.org/abs/1905.06731) (2019).
3. Li, W., et al.: Privacy-Preserving Federated Brain Tumour Segmentation. In: Suk, H.I., Liu, M., Yan, P., Lian, C. (eds) *Machine Learning in Medical Imaging. MLMI 2019. LNCS*, vol 11861, pp. 133–141. Springer, Cham (2019).
4. Sheller, M.J., et al.: Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. In: Crimi, A., Bakas, S., Kuijff, H., Keyvan, F., Reyes, M., van Walsum, T. (eds) *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2018. LNCS*, vol 11383, pp. 92–104. Springer, Cham (2019).
5. Wang, P. et al.: Automated Pancreas Segmentation Using Multi-institutional Collaborative Deep Learning. In: Albarqouni S. et al. (eds) *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning. DART 2020, DCL 2020. LNCS*, vol 12444, pp. 192–200. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-60548-3\\_19](https://doi.org/10.1007/978-3-030-60548-3_19)
6. Shen, C. et al.: Multi-task Federated Learning for Heterogeneous Pancreas Segmentation. In: Oyarzun Laura, C. et al. (eds) *Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning. DCL 2021, PPML 2021, LL-COVID19 2021, CLIP 2021. LNCS*, vol 12969, pp. 101–110. Springer, Cham (2021).
7. Xia, Y., et al.: Auto-FedAvg: Learnable Federated Averaging for Multi-Institutional Medical Image Segmentation. arXiv preprint [arXiv:2104.10195](https://arxiv.org/abs/2104.10195) (2021)
8. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-Efficient Learning of Deep Networks from Decentralized Data. In: *AISTATS* (2017).
9. Qureshi, T. et al. ‘Predicting Pancreatic Ductal Adenocarcinoma Using Artificial Intelligence Analysis of Pre-diagnostic Computed Tomography Images’. 1 Jan. 2022: 211–217. <https://doi.org/10.3233/CBM-210273>
10. Ashish Vaswani et al. “Attention is all you need.” In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS’17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.
11. Qureshi, T., et al: Morphology-guided deep learning framework for segmentation of pancreas in computed tomography images. *J. Med. Imag.* 9(2) 024002 (4 April 2022)

12. Nilsson, A., et al.: A Performance Evaluation of Federated Learning Algorithms. In: Proceedings of the Second Workshop on Distributed Infrastructures for Deep Learning (DIDL'18), pp. 1–8. Association for Computing Machinery, New York, NY, USA (2018).
13. Simpson, A. L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint [arXiv:1902.09063](https://arxiv.org/abs/1902.09063).
14. Holger R. Roth, Amal Farag, Evrim B. Turkbey, Le Lu, Jiamin Liu, and Ronald M. Summers. (2016). Data From Pancreas-CT. The Cancer Imaging Archive. <https://doi.org/10.7937/K9/TCIA.2016.tNB1kqBU>
15. Landman, BA, Xu, Z, Igelsias, JE, Styner, M, Langerak, TR, Klein, A: MICCAI multi-atlas labeling beyond the cranial vault—workshop and challenge, 2015.
16. The MONAI Consortium. (2020). Project MONAI. Zenodo.
17. Elman, J.L. (1990), Finding Structure in Time. *Cognitive Science*, 14: 179–211.
18. NVIDIA Clara Imaging. <https://developer.nvidia.com/clar-medical-imaging> (2022).
19. Kerfoot, E., Clough, J., Oksuz, I., Lee, J., King, A.P., Schnabel, J.A. (2019). Left-Ventricle Quantification Using Residual U-Net. In: , *et al.* *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges. STACOM 2018. Lecture Notes in Computer Science*, vol 11395. Springer, Cham.

# UNet-2022: Exploring Dynamics in Non-isomorphic Architecture



Jiansen Guo, Hong-Yu Zhou, Liansheng Wang, and Yizhou Yu

**Abstract** In this paper, we first analyze the differences between the weight allocation mechanisms of the self-attention and convolution. Based on this analysis, we propose to construct a parallel non-isomorphic block that takes the advantages of self-attention and convolution with simple parallelization. We name the resulting U-shape segmentation model as UNet-2022. In experiments, UNet-2022 obviously outperforms its counterparts in a range segmentation tasks, including abdominal multi-organ segmentation, automatic cardiac diagnosis, neural structures segmentation, and skin lesion segmentation, sometimes surpassing the best performing baseline by 4%. Specifically, UNet-2022 surpasses nnUNet, the most recognized segmentation model at present, by large margins. These phenomena indicate the potential of UNet-2022 to become the model of choice for medical image segmentation.

**Keywords** Medical image segmentation · Transformer

## 1 Introduction

Image segmentation has been among the fundamental tasks in medical image analysis. As the most widely adopted segmentation tools, UNet [1] and most of its series [2–4] were built upon DCNNs. With the prevalence of vision transformers in 2021, the medical imaging community started to incorporate the self-attention module into U-shape segmentation models for performance boosting [5–15]. The core behind these approaches is to construct non-isomorphic U-shape architecture by integrating self-attention with convolution. Although these methods achieved progress in different medical imaging tasks, most of them failed to provide an intuitive expla-

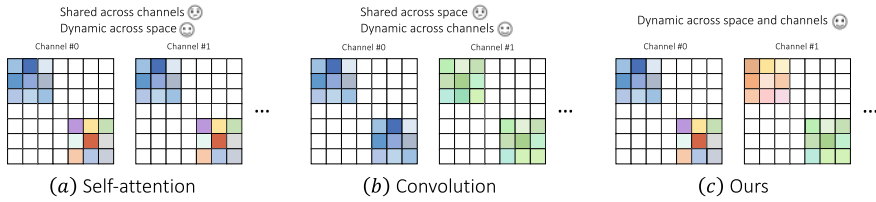
---

Jiansen Guo, Hong-Yu Zhou authors are contributed equally.

---

J. Guo · L. Wang (✉)  
School of Informatics, Xiamen University, Xiamen, China  
e-mail: [lswang@xmu.edu.cn](mailto:lswang@xmu.edu.cn)

H.-Y. Zhou · Y. Yu  
Department of Computer Science, The University of Hong Kong, Hong Kong, China



**Fig. 1** Illustration of the weight allocation mechanisms. Different colors denote different weights

nation for why this combination can be optimal. Accordingly, it is unclear how to better exploit the advantages of self-attention and convolution to build more optimal segmentation networks.

Let us briefly review the weight allocation mechanisms of self-attention and convolution, respectively. As is well-known, the key characteristic that led to the success of Transformers is the self-attention mechanism [16]. In vision transformers [17, 18], self-attention relates representations at different positions by employing a dynamic weight allocation mechanism. Thus, as shown in Fig. 1a self-attention, different positions have different weights while all channels at the same position share the same weight. On the other hand, DCNNs rely on extra learnable convolution kernels to aggregate spatial representations. As shown in Fig. 1b convolution, the same set of convolution kernel weights are shared across different spatial positions while dynamic weights are assigned to different channels.

Based on the above analysis, we see that self-attention and convolution maintain distinct but complementary characteristics. Based on this insight, we introduce a non-isomorphic block to include self-attention and convolution as two parallel modules. The proposed block comprises a novel weight allocation mechanism, which introduces dynamic to both space and channel dimensions (cf Fig. 1). In practice, we find this embarrassingly simple combination performs surprisingly well, outperforming previous state-of-the-art medical segmentation models by large margins in various segmentation tasks. Moreover, to reduce the risk of overfitting, we use depth-wise convolution (DWConv) for decreasing the number of weight parameters, which we empirically found performs slightly better than the naive convolution. The resulting UNet-2022 obviously outperforms nnUNet, currently the best generic medical image segmentation model, in a range of medical image segmentation tasks, including abdominal multi-organ segmentation, automatic cardiac diagnosis, neural structures segmentation, and skin lesion segmentation. For instance, UNet-2022 surpasses nnUNet by nearly 4% with a much smaller input size on multi-organ segmentation.

## 2 Related Work

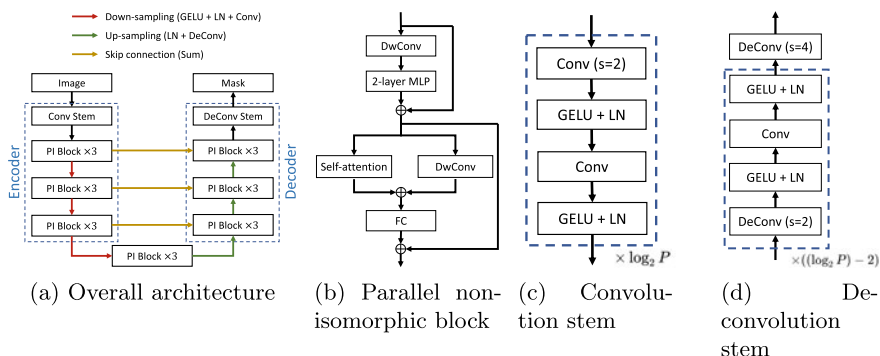
TransUNet [7], TransClaw UNet [6], and LeViT-UNet [13] inserted self-attention layers between the encoder and decoder of DCNNs to take advantage of captur-



ing long-range dependencies among a number of feature channels. Swin UNet [5] replaced the convolutional blocks with the Swin Transformer blocks [18] and built a U-shape segmentation model. DS-TransUNet [19] extended Swin UNet by introducing a fusion module for modeling long-range dependencies between features of different scales. Similar to Swin UNet, MISSFormer [8] built a hierarchical U-shape Transformer network that bridges all stages from the encoder to the decoder by mixing multi-scale information obtained by the hierarchical Transformer blocks. UNETR [20] adopted ViT [17] as the encoder network. In UNETR, feature maps from different layers of ViT with different resolutions are collected and sent to the convolutional decoder to capture the multi-scale information. nnFormer [14] utilized both local and global volume-based self-attention to build feature pyramids. MedT [21] proposed a gated axial attention layer which introduces a summational control mechanism in the self-attention. However, one problem of these hybrid segmentation models is that they did not provide an intuitive explanation of why the combination of self-attention and convolution can be beneficial. As a result, it is still unclear how to build an optimal combination of self-attention and convolution.

### 3 Methodology

We illustrate the detailed architecture of UNet-2022 in Fig. 2. As shown in Fig. 2a, the encoder of UNet-2022 consists of one convolution stem and three stages, where each stage involves three parallel non-isomorphic (PI) blocks. Symmetrically, the decoder also comprises three stages and one de-convolution stem. At each down-sampling/up-sampling step, we increase/decrease the number of channels and decrease/increase the spatial resolution of feature maps accordingly. Skip connections are used to bridge the gap between low-level details and high-level semantics.



**Fig. 2** Illustrations of UNet-2022. **DwConv**, **MLP**, **FC**, and **LN** stand for the depth-wise convolution, multi-layer perceptron, fully-connection layer, and layer normalization.  $P$  is a hyper-parameter, which varies based on the input resolution

Figure 2b describes the internal structural details of the parallel non-isomorphic block. We use depth-wise convolution (DwConv) to reduce the number of parameters. Self-attention and depth-wise convolution are parallelized to explore dynamics in spatial and channel dimensions, respectively, whose outputs are then added up and passed to a fully-connected (FC) layer. In the convolution stem, we stack multiple convolution layers to extract high-resolution feature maps, inspired by [14]. Similar operations are also applied in the de-convolution stem, where multiple de-convolution layers are stacked to produce the final segmentation mask.

### 3.1 Parallel Non-isomorphic Block

As aforementioned, self-attention and convolution emphasize dynamics in different dimensions, making them complementary to each other. Inspired by this finding, we propose to integrate their advantages in a non-isomorphic block via straightforward parallelization on self-attention and convolution.

Suppose  $\mathcal{F}^l \in \mathbb{R}^{H \times W \times C}$  denotes the input feature map to a parallel non-isomorphic block. As shown in Fig. 2b, we first pass the input feature map  $\mathcal{F}^l$  to a DwConv layer, after which a 2-layer multi-layer perceptron (MLP) is appended. The internal structures of the MLP layer are as follows: LayerNorm (LN)-FC-GELU-FC. We also add a residual connection to the output of MLP:

$$\hat{\mathcal{F}}^l = \text{MLP}(\text{DwConv}(\mathcal{F}^l)) + \mathcal{F}^l. \quad (1)$$

$\hat{\mathcal{F}}^l$  is then forwarded to two parallel layers, i.e., self-attention and DwConv, whose outputs are added up:

$$\tilde{\mathcal{F}}^l = \text{SA}(\hat{\mathcal{F}}^l) + \text{DwConv}(\hat{\mathcal{F}}^l), \quad (2)$$

where SA stands for the self-attention layer. Here we employed the window-based self-attention layer, proposed in [18], to improve the running efficiency. The kernel size of DwConv is 7. Finally,  $\tilde{\mathcal{F}}^l$  is fed to the last FC layer and another residual connection is added:

$$\overline{\mathcal{F}}^l = \text{FC}(\tilde{\mathcal{F}}^l) + \hat{\mathcal{F}}^l. \quad (3)$$

$\overline{\mathcal{F}}^l$  is the output of the PI block and will be passed to the following layer as the input feature map.

## 3.2 Convolution/De-convolution Stem

As displayed in Fig. 2c, the convolution stem consists of a number of stacked blocks (layers in the dashed box). To adapt to images with high resolutions, we stack more blocks in the convolution stem in order to reduce the memory cost and improve the computational efficiency. Specifically, the number of blocks, i.e.,  $\log_2 P$ , depends on the patch size  $P$  that we manually set. For instance, we set  $P$  to 4 for small and medium input resolutions, such as  $224 \times 224$  and  $320 \times 320$ . We increase  $P$  to 8 when the input resolution is high, such as  $512 \times 512$ . Compared to the patchify stem used in [17, 18], we empirically found our convolution stem could better capture the low-level information with equal receptive fields. Similar to the convolution stem, the architecture of the de-convolution stem also varies based on the resolution of input images.

## 4 Experiments

### 4.1 Dataset

**Multi-organ CT segmentation (Synapse).** Synapse<sup>1</sup> consists of 30 abdominal CT scans, where 13 organs were annotations. After pre-processing, we extract 3779 slices from all CT cases. Following instructions from [7], we split the whole dataset into training (18 scans, 2211 slices) and test (12 scans, 1568 slices) sets.

**Automated cardiac diagnosis (ACDC) [22].** The dataset is split into 70 samples for training (1290 slices), 10 samples for validation (196 slices), and 20 samples for testing (416 slices).

**Neural structures segmentation (EM).** EM (Electron Microscopy) dataset [23] contains 30 images and the size of each image is  $512 \times 512$ . The whole dataset is split into 24 samples for training, 3 samples for validation, and 3 samples for test.

**Skin lesion segmentation (ISIC-2016 and PH2).** The training dataset comes from the International Skin Imaging Collaboration at year 2016 (ISIC-2016), which contains 900 samples with lesion segmentations from dermoscopic images. Following [24, 25], we construct the test set using images from PH2 [26].

---

<sup>1</sup> <https://www.synapse.org/#!/synapse:syn3193805/wiki/217789>

**Table 1** Comparisons with 2D DCNN-based and hybrid segmentation models on multi-organ segmentation (Synapse)

Methods	Size	Average		Aorta	Gallbladder	Kidney (Left)	Kidney (Right)	Liver	Pancreas	Spleen	Stomach
		DSC $\uparrow$	HD95 $\downarrow$								
ViT [17] + CUP [7]	224	67.86	36.11	70.19	45.10	74.70	67.40	91.32	42.00	81.75	70.44
R50-ViT [17] + CUP [7]	224	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet [7]	224	77.48	31.69	87.23	63.16	81.87	77.02	94.08	55.86	85.08	75.62
TransUNet [7]	512	84.36	–	90.68	<b>71.99</b>	86.04	83.71	95.54	73.96	88.80	<u>84.20</u>
TransClaw UNet [6]	224	78.09	26.38	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55
TransClaw UNet [6]	512	80.39	–	90.00	56.86	83.27	76.21	95.06	67.76	91.16	82.82
SwinUNet [5]	224	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
SwinUNet $\nabla$ [5]	384	81.12	–	87.07	70.53	84.64	82.87	94.72	63.73	90.14	75.29
LeViT-UNet-384s [13]	224	78.53	16.84	87.33	62.23	84.61	80.25	93.11	59.07	88.86	72.76
MT-UNet [27]	224	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MISSFormer [8]	224	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	<u>91.92</u>	80.81
mnUNet [2]	512	82.36	24.74	90.96	65.57	81.92	78.36	95.96	69.36	91.12	85.60
UNet-2022	224	<u>84.98</u>	<u>16.70</u>	<b>92.10</b>	<u>69.63</u>	<u>88.40</u>	<u>83.93</u>	<u>96.02</u>	<u>75.50</u>	90.40	83.86
UNet-2022	320	<b>86.46</b>	<b>11.34</b>	<u>91.96</u>	69.40	<b>89.26</b>	<b>85.58</b>	<b>96.34</b>	<b>75.66</b>	<b>94.22</b>	<b>89.29</b>

The best results are bolded while the second best are underlined. **Size** denotes the input size

## 4.2 Implementation Details

All experiments are implemented on a single NVIDIA 2080ti GPU with 11 GB memory. We utilize both cross-entropy loss and dice loss and add them up like the Eq. 4 where the  $\lambda_1$  and  $\lambda_2$  are 1.2 and 0.8 for all datasets.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{DSC}} + \lambda_2 \mathcal{L}_{\text{CE}}. \quad (4)$$

**Inference details.** During the inference stage, UNet-2022 makes predictions following a sliding window manner. On Synapse, we set the step size of each sliding window to  $0.3 \times \text{CROP\_SIZE}$  and  $0.2 \times \text{CROP\_SIZE}$  for  $224 \times 224$  and  $320 \times 320$  input sizes, respectively. A smaller step size means that more overlapped patches participate in the voting of the mask prediction, leading to better segmentation performance. On the rest three datasets, the crop size is close to the full image size.

**Table 2** Comparisons on automatic cardiac diagnosis (ACDC)

Methods	Ave. DSC $\uparrow$	RV	Myo	LV
ViT [17] + CUP [7]	81.45	81.46	70.71	92.18
R50-ViT [17] + CUP [7]	87.57	86.07	81.88	94.75
TransUNet [7]	89.71	88.86	84.54	95.73
SwinUNet [5]	90.00	88.55	85.62	95.83
LeViT-UNet-384s [13]	90.32	89.55	87.64	93.76
MISSFormer [8]	90.86	89.55	88.04	94.99
MT-UNet [27]	90.43	86.64	89.04	95.62
nnUNet [2]	<u>92.32</u>	<u>90.39</u>	<u>90.53</u>	<u>96.05</u>
UNet-2022	<b>92.83</b>	<b>91.04</b>	<b>90.97</b>	<b>96.49</b>

The evaluation metric is DSC (%). The best results are bolded while the second best are underlined. The default input size is  $224 \times 224$  for all approaches

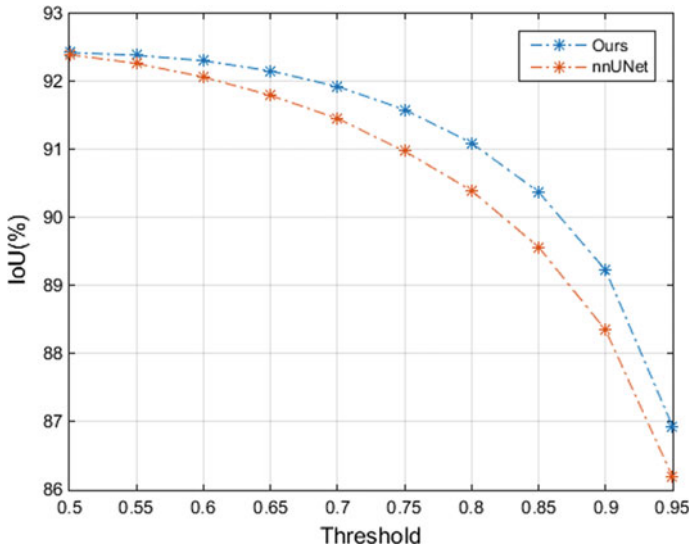
**Table 3** Comparisons on EM

Methods	mIoU $\uparrow$
UNet [28]	88.30
Wide UNet [4]	88.37
UNet+ [4]	88.89
UNet++ [4]	89.33
nnUNet [2]	<u>90.55</u>
UNet-2022	<b>91.05</b>

The evaluation metric is the mean IoU (mIoU). The best result is bolded while the second best is underlined

### 4.3 Comparisons on Abdominal Multi-organ Segmentation

Table 1 presents the segmentation performance on 8 organs on Synapse. When the input size is  $224 \times 224$ , we see that MISSFormer [8] achieves the highest average DSC while LeViT-UNet-384s [13] produces the lowest average HD95 among all baselines. In comparison, our UNet-2022 outperforms MISSFormer by about 3% in the average DSC. Despite average HD95 of our method is slightly better than that of LeViT-UNet-384s, our approach achieves a much higher average DSC, surpassing LeViT-UNet-384s by over 6%. Considering the trade-off between the running efficiency and task performance, we increase the input resolution to  $320 \times 320$  and observe large performance gains over  $224 \times 224$ . Comparing UNet-2022 with nnUNet [2], our method achieves impressive progress under both DSC and HD95 metrics. UNet-2022 outperforms nnUNet by 4% in the average DSC while dramatically reducing HD95 to 11.34 mm (14 mm reduction).



**Fig. 3** Comparisons of nnUNet and UNet-2022 on neural structures segmentation (EM). We present the computed IoU results at thresholds ranging from 0.5 to 0.95 with a step size of 0.05

#### 4.4 Comparisons on Automated Cardiac Diagnosis

In Table 2, we compare the segmentation performance of different models on ACDC. We fix the input resolution to  $224 \times 224$ . Somewhat surprisingly, we find that nnUNet obviously outperforms MISSFormer by nearly 1.5% on average while providing consistent performance gains on all three classes. Nonetheless, our UNet-2022 still outperforms nnUNet by about 0.5% in average. Moreover, UNet-2022 achieves consistent improvements on all three individual classes, demonstrating the potential of UNet-2022 to replace nnUNet.

#### 4.5 Comparisons on Neural Structures Segmentation

In this task, we follow [4] to calculate IoUs at thresholds ranging from 0.5 to 0.95 with a step size of 0.05. In Table 3, we present the segmentation performance of a range of UNet-like models. We see that nnUNet is the best performing baseline, outperforming the second best UNet++ by over 1%. Thus, we thoroughly compare UNet-2022 against nnUNet at different thresholds of IoU in Fig. 3. The default threshold is usually set as 0.5. As Fig. 3 displays, compared to nnUNet, UNet-2022 has obvious advantages in large thresholds. This phenomenon indicates that UNet-2022 is more advantageous in making high-confidence predictions.

**Table 4** Comparisons with 2D DCNN-based and hybrid segmentation models on skin lesion segmentation

Methods	DSC $\uparrow$	IoU $\uparrow$
SSLS [29]	78.3	68.1
MSCA [30]	81.5	72.3
FCN [31]	89.4	82.1
Bi et al. [32]	90.6	83.9
nnUNet [2]	91.6	85.1
Lee et al. [24]	91.8	84.3
BAT [25]	<u>92.1</u>	<u>85.8</u>
UNet-2022	<b>93.6</b>	<b>88.4</b>

The evaluation metric is DSC (%) and IoU(%). The best results are bolded while the second best are underlined

## 4.6 Comparisons on Skin Lesion Segmentation

Table 4 presents the comparisons of DCNN-based and hybrid models on skin lesion segmentation. The best performing baseline is BAT [25], which is a customized hybrid segmentation model that integrates the boundary-aware attention. BAT outperforms nnUNet by 0.5 and 0.7% in DSC and IoU, respectively.

Comparing UNet-2022 with BAT, we see that UNet-2022 achieves dramatic improvements in both DSC and IoU. For instance, UNet-2022 outperforms BAT by 2.6% in IoU while BAT only surpasses nnUNet by 0.7%. Considering IoU is a stricter metric than DSC, we believe the 2.6% improvement is sufficient enough to demonstrate the strengths of UNet-2022 over BAT and nnUNet.

## 4.7 Ablation Studies of Modules and Strategies

**Impact of the PI block.** Table 5 presents the comparisons of different building blocks, including blocks used in Swin Transformer [18], ConvNeXt [33], and our UNet-2022. We see that the CNX block performs slightly better than the ST block in average. Nonetheless, our PI block obviously surpasses the CNX block by 1.5% in the average DSC while largely improving HD95 by approximate 5 mm. The underlying reason is that the latter two blocks only use isomorphic operations. This characteristic makes them lack the ability to capture dynamics across different dimensions.

**Influences of dynamics across the space and channels.** In Table 6, we remove the self-attention layer from the PI block. As a result, the resulting building blocks fail to explore dynamics across different spatial positions, as they only contain convolution operations. This failure can be verified by the observable task performance drop in the second row of Table 6.

**Table 5** Ablations on building blocks

Methods	Average	
	DSC $\uparrow$	HD95 $\downarrow$
ST [18]	84.42	20.74
CNX [33]	84.96	16.60
Our PI	86.46	11.34

**Table 6** Influence of the self-attention (SA)

Methods	Average	
	DSC $\uparrow$	HD95 $\downarrow$
UNet-2022	86.46	11.34
• SA	85.20	14.96

**Table 7** Influence of the pre-training

Methods	Average	
	DSC $\uparrow$	HD95 $\downarrow$
UNet-2022	86.46	11.34
• Pre-training	84.87	15.04
ConvNeXt	84.96	16.60
• Pre-training	84.32	18.02

**Table 8** Impact of the inference step size

Step size	Average	
	DSC $\uparrow$	HD95 $\downarrow$
$0.5 \times \text{CROP SIZE}$	86.27	13.95
$0.2 \times \text{CROP SIZE}$	86.46	11.34

**Impact of ImageNet-based pre-training.** We use ImageNet-based pre-training to boost the segmentation performance. We replace the default encoder in UNet-2022 with the recently proposed ConvNeXt [33], and compare the modified ConvNeXt-based UNet-2022 with our proposed version on Synapse. As shown in Table 7, we see that ImageNet-based pre-training plays a vital role in both UNets, providing observable performance gains over training from scratch.

**Impact of the inference step size.** As aforementioned, we adjust the inference step size when sampling sliding windows on Synapse. Here, we present the impact of the step size in Table 8. When we set the step size to  $0.2 \times \text{CROP SIZE}$ , we find that it performs better than  $0.5 \times \text{CROP SIZE}$ , bringing about 2.6 mm improvement in HD95.



## 5 Conclusion

We build a non-isomorphic block by parallelizing self-attention and convolution operations. The resulting UNet-2022 achieves quite competitive performance in a range of medical image segmentation tasks. In the future, we will investigate how to appropriately incorporate self-supervised learning [34, 35] into UNet-2022 as we found pre-training plays a vital role in medical image segmentation.

## References

1. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
2. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* 18(2), 203–211 (2021)
3. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Transactions on Medical Imaging* 37(12), 2663–2674 (2018)
4. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging* 39(6), 1856–1867 (2019)
5. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint [arXiv:2105.05537](https://arxiv.org/abs/2105.05537) (2021)
6. Chang, Y., Menghan, H., Guangtao, Z., Xiao-Ping, Z.: Transclaw U-Net: Claw u-net with transformers for medical image segmentation. arXiv preprint [arXiv:2107.05188](https://arxiv.org/abs/2107.05188) (2021)
7. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint [arXiv:2102.04306](https://arxiv.org/abs/2102.04306) (2021)
8. Huang, X., Deng, Z., Li, D., Yuan, X.: MISSFormer: An effective medical image segmentation transformer. arXiv preprint [arXiv:2109.07162](https://arxiv.org/abs/2109.07162) (2021)
9. Liu, W., Tian, T., Xu, W., Yang, H., Pan, X.: PHTrans: Parallely aggregating global and local representations for medical image segmentation. arXiv preprint [arXiv:2203.04568](https://arxiv.org/abs/2203.04568) (2022)
10. Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M.: A volumetric transformer for accurate 3d tumor segmentation. arXiv preprint [arXiv:2111.13300](https://arxiv.org/abs/2111.13300) (2021)
11. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: TransBTS: Multimodal brain tumor segmentation using transformer. In: MICCAI. pp. 109–119. Springer (2021)
12. Xie, Y., Zhang, J., Shen, C., Xia, Y.: CoTr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: MICCAI. pp. 171–180. Springer (2021)
13. Xu, G., Wu, X., Zhang, X., He, X.: LeViT-UNet: Make faster encoders with transformer for medical image segmentation. arXiv preprint [arXiv:2107.08623](https://arxiv.org/abs/2107.08623) (2021)
14. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnFormer: Interleaved transformer for volumetric segmentation. arXiv preprint [arXiv:2109.03201](https://arxiv.org/abs/2109.03201) (2021)
15. Zhou, H.Y., Lu, C., Yang, S., Yu, Y.: Convnets vs. transformers: Whose visual representations are more transferable? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2230–2238 (2021)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *NeurIPS* 30 (2017)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)

18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
19. Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., Zhang, D.: DS-TransUNet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement* (2022)
20. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3d medical image segmentation. In: WACV. pp. 574–584 (2022)
21. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: MICCAI. pp. 36–46. Springer (2021)
22. Bernard, O., Lalonde, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging* 37(11), 2514–2525 (2018)
23. Cardona, A., Saalfeld, S., Preibisch, S., Schmid, B., Cheng, A., Pulokas, J., Tomancak, P., Hartenstein, V.: An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. *PLoS Biology* 8(10), e1000502 (2010)
24. Lee, H.J., Kim, J.U., Lee, S., Kim, H.G., Ro, Y.M.: Structure boundary preserving segmentation for medical image with ambiguous boundary. In: CVPR. pp. 4817–4826 (2020)
25. Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., Qin, J.: Boundary-aware transformers for skin lesion segmentation. In: MICCAI. pp. 206–216. Springer (2021)
26. Barata, C., Ruela, M., Francisco, M., Mendonça, T., Marques, J.S.: Two systems for the detection of melanomas in dermoscopy images using texture and color features. *IEEE Systems Journal* 8(3), 965–979 (2013)
27. Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.H., Chen, Y.W., Tong, R.: Mixed transformer u-net for medical image segmentation. *arXiv preprint [arXiv:2111.04734](https://arxiv.org/abs/2111.04734)* (2021)
28. Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., et al.: U-net: Deep learning for cell counting, detection, and morphometry. *Nature Method* 16(1), 67–70 (2019)
29. Ahn, E., Bi, L., Jung, Y.H., Kim, J., Li, C., Fulham, M., Feng, D.D.: Automated saliency-based lesion segmentation in dermoscopic images. In: EMBS. pp. 3009–3012. IEEE (2015)
30. Bi, L., Kim, J., Ahn, E., Feng, D., Fulham, M.: Automated skin lesion segmentation via image-wise supervised learning and multi-scale superpixel based cellular automata. In: ISBI. pp. 1059–1062. IEEE (2016)
31. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
32. Bi, L., Kim, J., Ahn, E., Kumar, A., Fulham, M., Feng, D.: Dermoscopic image segmentation via multistage fully convolutional networks. *IEEE Transactions on Biomedical Engineering* 64(9), 2065–2074 (2017)
33. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. *arXiv preprint [arXiv:2201.03545](https://arxiv.org/abs/2201.03545)* (2022)
34. Zhou, H.Y., Chen, X., Zhang, Y., Luo, R., Wang, L., Yu, Y.: Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence* 4(1), 32–40 (2022)
35. Zhou, H.Y., Lu, C., Yang, S., Han, X., Yu, Y.: Preservation learning improves self-supervised medical image models by reconstructing diverse contexts. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3499–3509 (2021)

# Hybrid-Fusion Transformer for Multisequence MRI



Jihoon Cho  and Jinah Park 

**Abstract** Medical segmentation has grown exponentially through the advent of a fully convolutional network (FCN), and we have now reached a turning point through the success of Transformer. However, the different characteristics of the modality have not been fully integrated into Transformer for medical segmentation. In this work, we propose the novel hybrid fusion Transformer (HFTrans) for multisequence MRI image segmentation. We take advantage of the differences among multimodal MRI sequences and utilize the Transformer layers to integrate the features extracted from each modality as well as the features of the early fused modalities. We validate the effectiveness of our hybrid-fusion method in three-dimensional (3D) medical segmentation. Experiments on two public datasets, BraTS2020 and MRBrainS18, show that the proposed method outperforms previous state-of-the-art methods on the task of brain tumor segmentation and brain structure segmentation.

**Keywords** Transformer · Multi-modality · 3D Medical image segmentation

## 1 Introduction

Magnetic resonance imaging (MRI) is widely used in the detection, diagnosis, and treatment planning of diseases in the human body, including the brain, spinal cord, prostate, and knee. Depending on the target organ and purpose, there are several types of MRI protocols consisting of many sequences [18]. Each MRI sequence has shown various characteristics, especially the signal of different tissues such as fluid, muscle, and fat. In addition, some sequences represent functional information beyond the anatomical structure [21]. Considering that a valuable feature varies by sequence type, a combination of sequences gives better results than unimodal processing in the presence of diseases [17] and lesion segmentation [4].

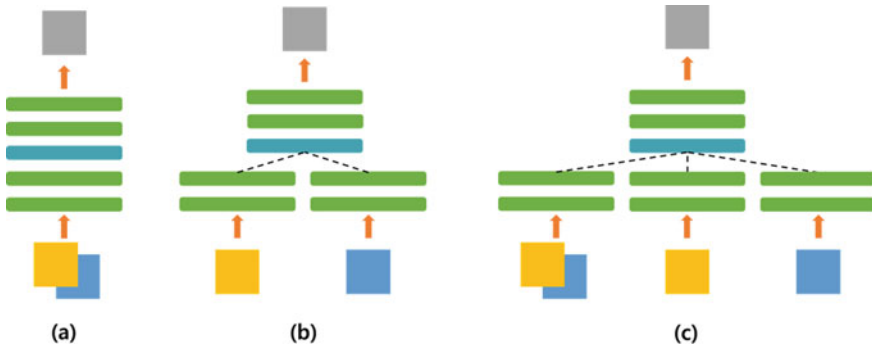
In recent years, Convolutional Neural Networks (CNN) have been successful in various computer vision tasks. U-Net [16] adopts the concept of a FCN [14] with a

---

J. Cho · J. Park (✉)

KAIST, 291 Daehak-Ro, Yuseong-Gu, Daejeon 34141, Republic of Korea

e-mail: [jinahpark@kaist.ac.kr](mailto:jinahpark@kaist.ac.kr)

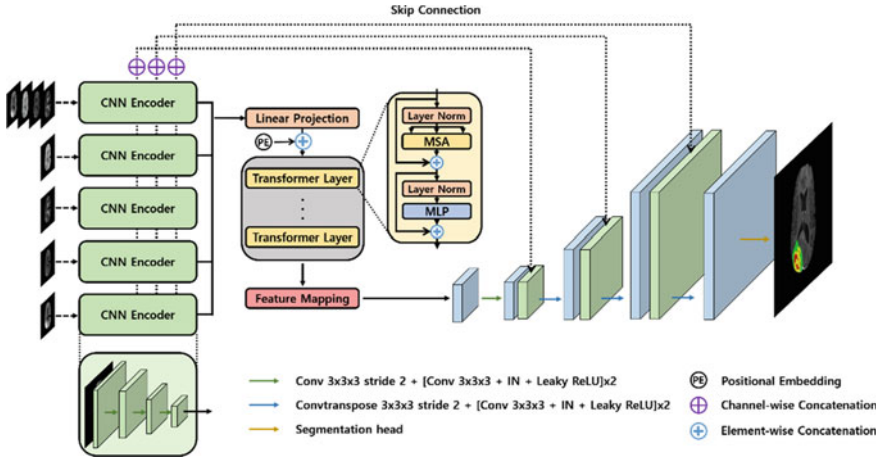


**Fig. 1** Three types of fusion methods. **a** early fusion **b** middle fusion **c** our hybrid fusion

relatively shallow structure and balancing feature representation and locality through the skip connection. Following the modifications suitable for 3D medical image segmentation, U-net has become the de facto standard even for multimodal MRI volumes with the early fusion of simple multichannel input [9, 10] as shown in Fig. 1a. More recently, Vision Transformer (ViT) [3], inspired by the tremendous success of Transformer [19], become a new solution to the limited receptive field of CNN with the global self-attention mechanism. For 3D medical segmentation, UNETR [6] and Swin-UNETR [5] have proposed Transformer networks with CNN layers on the decoder, and TransBTS [20] have constructed with CNN encoder and decoder with the bottom Transformer layer. However, all of these Transformer networks have the disadvantage of treating multiple MRIs as a multichannel input.

Similar to multisequence MRI, RGB-D images consist of multiple modalities that have the same spatial information: color image and depth image. However, considering the sharable and specific features between color and depth images [7], particular encoding (as shown in Fig. 1b is used for each modality in many tasks including semantic segmentation [11]. This approach has been used even for recent work of Transformer-based methods [12, 13]. From the effort to consider the multimodalities of RGB-D images, we find that the adoption of the middle fusion approach for MRI sequences can benefit from different modality characteristics.

In this work, inspired by the processing of multimodal RGB-D images and the long-range visual dependence from ViT, we propose the Hybrid-Fusion Transformer (HFTrans) for multisequence MRI images. The proposed HFTrans is constructed with the hybrid fusion approach to take advantage of both early fusion and middle fusion, as shown in Fig. 1c, and consists of multiple CNN encoders and the Transformer encoder. Each encoder extracts a local context feature representation for each modality, including the early fused modalities, and they are integrated in the Transformer encoder. The feature embedding from the Transformer encoder is progressively up-sampled with the spatial information from encoders via skip-connection, and finally predicts segmentation maps of the original resolution. In experiments on the Brain Tumor Segmentation 2020 dataset (BraTS2020) [2] and the MR Brain



**Fig. 2** Overview of HFTrans network for BraTS2020 dataset. Hybrid fusion of four MRI sequences is performed in the CNN Encoders and the Transformer Encoder. The encoded representation output from the Transformer encoder is progressively upsampled with skip connection to predict the final segmentation maps

Segmentation 2018 dataset (MRBrainS18),<sup>1</sup> we validate the effectiveness of our method in multisequence MRI segmentation. HFTrans achieves remarkable performance on both public challenge datasets. We also conduct further experiments on encoder compositions, which show that our hybrid fusion method works well without human heuristics by using simple encoders for each multisequence MRI image.

## 2 Method

An overview of HFTrans is presented in Fig. 2. Although we accept the early fusion encoding, the hybrid fusion method is applied by constructing additional encoders for each modality.

### 2.1 Hybrid Fusion from CNN Encoders

Considering the high computational cost of Transformer for high-resolution 3D images and the inductive bias of the convolutional layer, we propose to construct the convolutional layers to make a rich local context feature representation. To bring benefits from different modality characteristics, each modality is processed in individual encoders. Features are embedded into 1D sequences and then perform

<sup>1</sup> <https://mrbrains18.isi.uu.nl/>

self-attention between feature embedding in the Transformer layer. In addition, the encoding of early fused modalities is also utilized, taking into account the ability to extract an apparent powerful representation. Hybrid fusion between powerful representation from the entire encoding and modality-specific representation from separate encoding can exploit the advantages of both methods.

For the 3D MRI input consisting of  $N$  MRI sequences  $x_i (i = 1, 2, \dots, N) \in \mathbb{R}^{1 \times W \times H \times D}$  with resolution  $(W, H, D)$ , we use  $N + 1$  feature representations for hybrid fusion.  $N$  features are extracted individually from each MRI sequence, and the representation of early fusion is encoded by all  $N$  sequences  $x \in \mathbb{R}^{N \times W \times H \times D}$ . The encoders have the same structure consisting of stacking the convolutional layers  $3 \times 3 \times 3$  and stride-convolutional layers consecutively. Then, the high-level feature representations  $f_j (j = 1, 2, \dots, N + 1) \in \mathbb{R}^{K \times \frac{w}{8} \times \frac{h}{8} \times \frac{d}{8}}$  are projected linearly, but the computational complexity of the Transformer layer is increased quadratic based on the number of 1D sequences. Therefore, we apply the  $2 \times 2 \times 2$  patch embedding projection to the features extracted from CNN. Subsequently, we get the input embedding  $z_0 \in \mathbb{R}^{C \times M (= N \times \frac{w}{16} \times \frac{h}{16} \times \frac{d}{8})}$  with the channel dimension  $C$ . To preserve location information of flattened sequences, we add a learnable positional embedding  $E_{pos} \in \mathbb{R}^{C \times M}$  as

$$z_0 = W \times f + E_{pos} \quad (1)$$

where  $W$  is the 1D projector with  $2 \times 2 \times 2$  patch. After the feature embedding, we conduct self-attention using a standard Transformer encoder consisting of  $L$  Transformer layers. The  $l$ -th Transformer layer is operated as follows,

$$z_l^* = MSA(LN(z_{l-1})) + z_{l-1} \quad (2)$$

$$z_l = MLP(LN(z_l^*)) + z_l^* \quad (3)$$

where  $MSA$  denotes multihead self-attention,  $MLP$  is multilayer perceptron, and  $LN$  refers to layer normalization.

## 2.2 CNN Decoder and Loss Function

The output sequences of the Transformer encoder are reshaped in the 4D feature maps to generate voxel-wise semantic segmentation results. The reshaped feature maps  $d_j \in \mathbb{R}^{C \times \frac{w}{16} \times \frac{h}{16} \times \frac{d}{16}}$  are concatenated channel-wise and upsampled by a factor of 2 to shape the original feature size of  $f_j$  before linear projection. After feature mapping, the feature representation  $d \in \mathbb{R}^{K \times \frac{w}{8} \times \frac{h}{8} \times \frac{d}{8}}$  is progressively fed into the deconvolution of stride 2 and the convolutional layers  $3 \times 3 \times 3$ . During deconvolution, we aggregate the encoding features of multiple CNN encoders via a skip connection. This process is repeated up to the feature representation reaching the original input resolution,

and the final semantic segmentation is generated through the convolutional layer  $1 \times 1 \times 1$  with a softmax activation function. We use both Dice loss and cross-entropy loss together as an objective function.

### 3 Experiments

#### 3.1 Datasets

We use two publicly available 3D medical segmentation datasets consisting of multimodal MRI images: BraTS2020 and MRBrainS18.

**BraTS2020:** BraTS2020 [2] is a patient’s brain MRI dataset labeled with three tumor sub-regions, peritumoral edematous tissue, enhancing tumor, and necrotic tumor core. The dataset contains 369 training sets acquired from several institutions with various protocols and scanners. Each MRI scan consists of four sequences: T1-weighted (T1), T2-weighted (T2), post-contrast T1-weighted (T1ce), and T2 fluid-attenuated inversion recovery (FLAIR). They were provided after preprocessing of the co-registration and skull stripping, and we additionally perform z-score normalization to brain regions except the masked background area with zero intensity. All MRI sequences have the same voxel size of  $240 \times 240 \times 155$  with 1 mm isotropic voxel spacing.

**MRBrainS18:** For the whole brain segmentation, we use the MRBrainS18 dataset that includes both brain structure and pathological abnormalities. The dataset consists of 30 subjects acquired on a 3 T scanner from various patients, including dementia, diabetes, and Alzheimer’s. Multimodal MRI scans consist of aligned sequences of T1, T1 inversion recovery sequence (T1-IR), and FLAIR. All scans have a  $0.958 \text{ mm} \times 0.958 \text{ mm} \times 3 \text{ mm}$  voxel spacing with  $240 \times 240 \times 48$  voxel size. We perform a sevenfold cross-validation for 7 training set and use 8 labels in the evaluation, which are gray matter, basal ganglia, white matter, white matter lesion, CSF, ventricles, cerebellum, and brain stem.

#### 3.2 Quantitative Results

We perform experiments on BraTS2020 and MRBrainS18 datasets by comparing our HFTrans with five previous state-of-the-art: (1) U-Net [16]; (2) ResUNet [22]; (3) AttnUNet [15]; (4) nnU-Net [8]; (5) TransBTS [20], which is the Transformer-based network with an early fusion approach. We perform a five-fold cross-validation on the BraTS2020 dataset for all methods. As shown in Table 1, HFTrans achieves Dice scores of 82.81%, 84.66%, 90.82% and HD95 of 26.42 mm, 6.98 mm, 2.57 mm on ET, TC, WT, which are higher results than the other methods except HD95 of TC.

**Table 1** Cross-validation results on the BraTS2020 dataset. ET, TC, and WT denote enhancing tumor, tumor core, and whole tumor

Models	Dice score (%) $\uparrow$				HD95 (mm) $\downarrow$			
	ET	TC	WT	Avg	ET	TC	WT	Avg
U-Net [16]	80.79	80.40	88.67	83.29	32.20	17.13	4.15	17.83
ResUNet [22]	80.31	78.45	88.79	82.52	29.05	16.11	4.70	16.62
AttnUNet [15]	80.73	79.30	88.54	82.86	31.24	23.86	11.30	22.13
nnU-Net [8]	82.28	84.18	90.56	85.67	32.20	<b>5.03</b>	2.68	13.30
TransBTS [20]	81.39	80.70	90.16	84.08	30.59	15.17	7.68	17.81
HFTrans	<b>82.81</b>	<b>84.66</b>	<b>90.82</b>	<b>86.10</b>	<b>26.42</b>	6.98	<b>2.57</b>	<b>11.99</b>
HFTrans*	82.52	84.59	90.40	85.84	29.79	5.61	3.98	13.13

Compared to U-Net [16], ResUNet [22], AttnUNet [15], TransBTS [20], and nnU-Net [8], our proposed method outperforms them by 2.81%, 3.58%, 3.24%, 2.02% and 0.53% in terms of average Dice score and 5.84 mm, 4.63 mm, 10.14 mm, 5.82 mm, and 1.31 mm in terms of average HD95, respectively. HFTrans\*, the hybrid fusion variant model that consists of modality exception encoders instead of each modality encoder (described in Table 3), also outperforms the previous methods.

The results evaluated on MRBrainS18 are reported in Table 2. HFTrans achieves Dice score 84.81%, HD95 3.25 mm, and volume similarity 94.12%, which outperforms the result of nnU-Net [8] and TransBTS [20] by 2.16% and 1.44% in terms of Dice score, 3.29 mm and 2.01 mm in terms of HD95, and 1.49% and 1.08% in terms of volume similarity. It is also comparable to U-Net [16], ResUNet [22], and AttnUNet [15]. Comparing the model complexity, U-Net, ResUNet, AttnUNet, and our HFTrans have 90.30 M, 37.72 M, and 25.78 M, 65.17 M parameters and 266.91G, 498.53G, 329.54G, and 140.39G FLOPs, respectively. Despite the relatively small model complexity, HFTrans shows significantly better performance, especially in brain stem segmentation, by bridging high-level global context information with low-level local details.

### 3.3 Qualitative Results

Qualitative comparisons on brain tumor segmentation are presented in Fig. 3. Our hybrid fusion method HFTrans shows fine-grained segmentation of brain tumors, while the pure CNN-based method nnU-Net tends to over-segment and the CNN-Transformer method TransBTS tends to under-segment, which are evident in rows 1 and 3. This indicates that hybrid fusion captures both powerful spatial context and long-range dependency. In Fig. 4, we present qualitative segmentation comparisons for brain structure segmentation in the MRBrainS18 dataset. HFTrans exhibits detailed segmentation of the whole brain structure. In particular, our method shows



**Table 2** Cross-validation results on the MRBrainS18 dataset

Models	Dice score (%) $\uparrow$								
	GM	BG	WM	WML	CSF	Vent	Cereb	BS	Avg
U-Net [16]	<b>84.79</b>	83.89	86.24	<b>64.86</b>	<b>82.82</b>	93.60	<b>92.62</b>	88.59	84.67
ResUNet [22]	84.23	83.51	86.06	64.27	82.29	93.29	92.17	88.88	84.34
AttnUNet [15]	84.68	83.17	86.51	63.44	82.54	<b>93.69</b>	92.46	89.50	84.50
nnU-Net [8]	82.60	80.99	85.49	60.59	79.87	92.59	91.00	88.09	82.65
TransBTS [20]	83.07	83.71	85.78	60.59	80.43	92.57	92.35	88.47	83.37
HFTrans	84.71	83.74	<b>86.99</b>	64.03	82.35	93.50	92.32	<b>90.85</b>	<b>84.81</b>
HFTrans*	84.33	<b>84.17</b>	86.80	63.80	82.46	93.59	91.40	90.60	84.64
Models	HD95 (mm) $\downarrow$								
	GM	BG	WM	WML	CSF	Vent	Cereb	BS	Avg
U-Net [16]	<b>0.96</b>	3.07	<b>1.15</b>	10.83	<b>1.98</b>	<b>1.36</b>	1.36	3.85	3.30
ResUNet [22]	1.01	3.10	1.51	10.76	2.04	1.53	1.53	3.95	3.37
AttnUNet [15]	<b>0.96</b>	3.11	1.48	10.95	<b>1.98</b>	1.48	1.48	3.46	3.31
nnU-Net [8]	1.52	3.86	1.96	12.38	2.44	1.78	1.78	25.26	6.54
TransBTS [20]	1.18	3.02	1.84	12.60	2.41	2.12	2.12	16.02	5.26
HFTrans	1.15	<b>2.85</b>	1.48	<b>10.36</b>	2.08	2.41	<b>2.79</b>	<b>2.90</b>	<b>3.25</b>
HFTrans*	1.07	3.05	1.47	11.15	<b>1.98</b>	<b>1.36</b>	3.23	3.21	3.32
Models	Volume similarity (%) $\uparrow$								
	GM	BG	WM	WML	CSF	Vent	Cereb	BS	Avg
U-Net [16]	95.16	94.91	94.42	<b>82.96</b>	94.95	96.91	96.24	93.63	93.49
ResUNet [22]	95.20	94.06	94.43	80.61	93.53	94.44	95.91	95.04	93.28
AttnUNet [15]	95.48	93.68	95.29	79.39	<b>95.09</b>	96.63	95.96	94.27	93.30
nnU-Net [8]	95.37	93.47	<b>96.12</b>	76.55	94.38	97.46	95.19	93.01	92.63
TransBTS [20]	95.73	94.81	96.02	73.17	95.03	<b>97.53</b>	<b>97.49</b>	94.67	93.04
HFTrans	<b>95.74</b>	94.16	96.04	80.59	94.97	96.92	96.63	96.03	<b>94.12</b>
HFTrans*	95.17	<b>95.49</b>	95.61	78.30	94.53	97.14	95.38	<b>96.13</b>	93.55

Note GM: gray matter, BG: basal ganglia, WM: white matter, WML: white matter lesions, CSF: cerebrospinal fluid, Vent: ventricles, Cereb: cerebellum, BS: brain stem

superior performance with a detailed boundary in brain stem segmentation, and the effectiveness of the hybrid fusion method is demonstrated.

## 4 Discussion

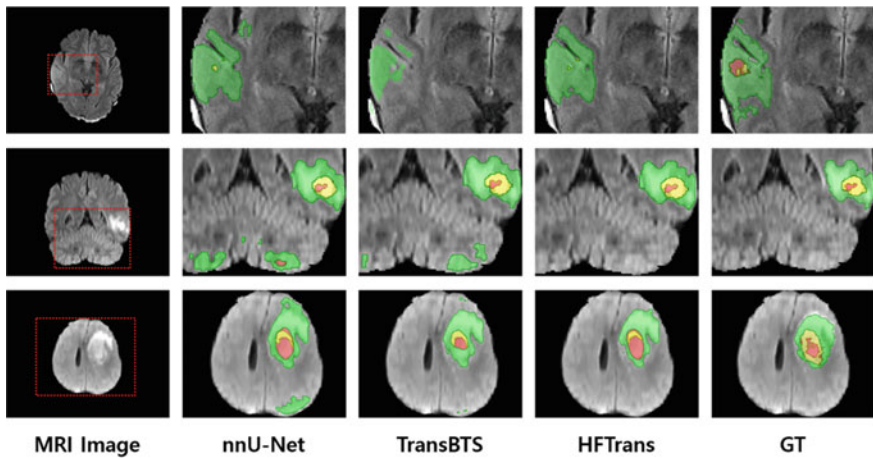
We evaluate the effectiveness of our encoder composition by comparing the early fusion approach, which takes all modality as input, the middle fusion approach of

**Table 3** Results for different variants of encoder composition

Encoder composition	Dice (%)	HD95 (mm)
T1/T2/T1ce/FLAIR (Middle Fusion)	82.40	28.01
All (Early fusion)	83.06	25.56
All/T1ce	82.62	27.35
All/FLAIR	83.02	24.09
All/T1ce/FLAIR	83.17	25.62
All/T1 + T1ce/FLAIR	82.58	27.29
All/T1/T2/T1ce/FLAIR (HFTrans)	<b>83.52</b>	24.07
All/T1*/T2*/T1ce*/FLAIR* (HFTrans*)	83.28	<b>22.63</b>

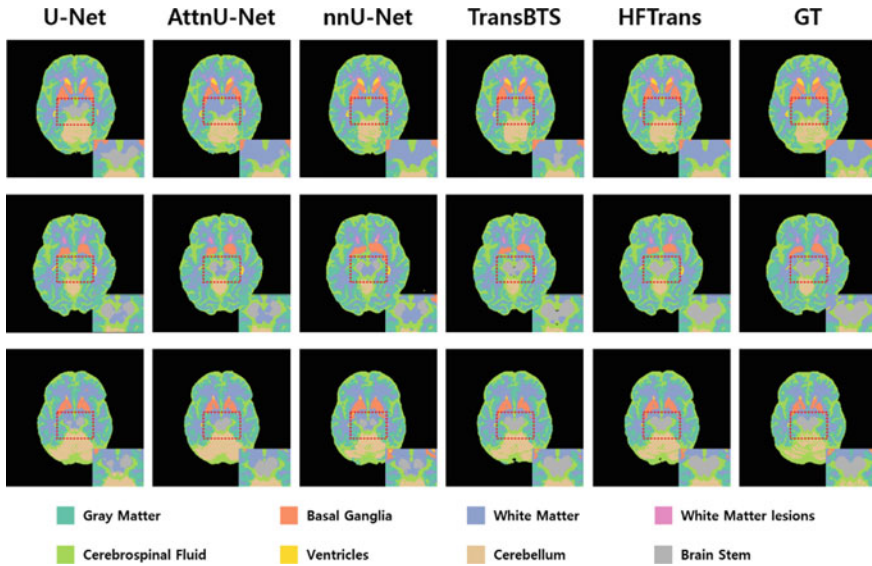
T1\* denotes the three-channel input of T2, T1ce, and FLAIR except for T1. T2\*, T1ce\*, and FLAIR\* have the same approach as T1\*

We compare the early fusion method, the middle fusion method, and our hybrid fusion methods including the additional human heuristics



**Fig. 3** Qualitative comparison of brain tumor segmentation on the BraTS2020 dataset. The enhancing tumor (ET) is depicted in the yellow region, and the tumor core (TC) is represented as a union of red and yellow regions. The whole tumor (WT) contains a colored region of green, red, and yellow

individual feature extraction from modalities, and considering the human heuristics of the annotation protocol [1], that the appearance of a brain tumor is typically depicted as a hyperintense signal in T1ce and FLAIR. As shown in Table 3, the middle fusion approach shows the worst results of Dice score 82.40% and HD95 28.01 mm, failing to get the benefit of each modality. The early fusion approach shows the better results of 83.06% and 25.56 mm in terms of Dice score and HD95. Several results of different human heuristic approaches improve HD95 of 3.92 mm when using the early fusion encoder and additional FLARE encoder, and improve Dice score of 0.77% when



**Fig. 4** Qualitative comparison of brain structure segmentation on the MRBrainS18 dataset. The brain stem region (gray) is zoomed-in

using the early fusion encoder, FLARE and T1ce encoders. However, they do not produce an improvement for both the Dice score and HD95 at the same time compared to the early fusion approach. The encoder compositions of our proposed method HFTrans, taking advantage of early fusion and middle fusion, improve performance by 1.12% and 3.94 mm on Dice Score and HD95 without human heuristics. In addition, the variant of our method, HFTrans\*, also shows improvements in both metrics, especially with a remarkable HD95 result of 22.63 mm.

## 5 Conclusion

This paper introduces a novel Transformer-based segmentation framework for multi-sequence MRI. The proposed hybrid fusion method inherits the advantages of the early fusion approach with the powerful locality of 3D CNN and the middle fusion approach with the global consistency of Transformer. Experiments on different volumetric segmentation datasets, BraTS2020 and MRBrainS18, validate the effectiveness of our method. The proposed method could serve as the basis for a Transformer-based segmentation network for multimodal medical images. As a future work, we plan to explore the Transformer-based fusion method with a focus on the computational efficiency.

**Acknowledgements** This work was supported by Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea government (MOTIE) (20201510300280, Development of a remote dismantling training system with force-torque responding virtual nuclear power plant).

## References

1. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint [arXiv:2107.02314](https://arxiv.org/abs/2107.02314) (2021)
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data* 4(1), 1–13 (2017)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
4. Garcia-Lorenzo, D., Prima, S., Arnold, D.L., Collins, D.L., Barillot, C.: Trimmed-likelihood estimation for focal lesions and tissue segmentation in multisequence mri for multiple sclerosis. *IEEE transactions on medical imaging* 30(8), 1455–1467 (2011)
5. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. arXiv preprint [arXiv:2201.01266](https://arxiv.org/abs/2201.01266) (2022)
6. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D.: Unetr: Transformers for 3d medical image segmentation (2021), <https://arxiv.org/abs/2103.10504>
7. Hu, J.F., Zheng, W.S., Lai, J., Zhang, J.: Jointly learning heterogeneous features for rgb-d activity recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 5344–5352 (2015)
8. Isensee, F., Jäger, P.F., Full, P.M., Vollmuth, P., Maier-Hein, K.H.: nnu-net for brain tumor segmentation. In: *International MICCAI Brainlesion Workshop*. pp. 118–132. Springer (2020)
9. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint [arXiv:1809.10486](https://arxiv.org/abs/1809.10486) (2018)
10. Kayalibay, B., Jensen, G., van der Smagt, P.: Cnn-based segmentation of medical imaging data. arXiv preprint [arXiv:1701.03056](https://arxiv.org/abs/1701.03056) (2017)
11. Lin, D., Chen, G., Cohen-Or, D., Heng, P.A., Huang, H.: Cascaded feature network for semantic segmentation of rgb-d images. In: *Proceedings of the IEEE international conference on computer vision*. pp. 1311–1319 (2017)
12. Liu, Z., Tan, Y., He, Q., Xiao, Y.: Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* 32(7), 4486–4497 (2021)
13. Liu, Z., Wang, Y., Tu, Z., Xiao, Y., Tang, B.: Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network. In: *Proceedings of the 29th ACM international conference on multimedia*. pp. 4481–4490 (2021)
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3431–3440 (2015)
15. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint [arXiv:1804.03999](https://arxiv.org/abs/1804.03999) (2018)

16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. CoRR abs/1505.04597 (2015), <http://arxiv.org/abs/1505.04597>
17. Schouten, T.M., Koini, M., De Vos, F., Seiler, S., Van Der Grond, J., Lechner, A., Hafkemeijer, A., M'oller, C., Schmidt, R., De Rooij, M., et al.: Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate alzheimer's disease. *NeuroImage: Clinical* 11, 46–51 (2016)
18. Traboulsee, A., Simon, J., Stone, L., Fisher, E., Jones, D., Malhotra, A., Newsome, S., Oh, J., Reich, D., Richert, N., et al.: Revised recommendations of the consortium of ms centers task force for a standardized mri protocol and clinical guidelines for the diagnosis and follow-up of multiple sclerosis. *American Journal of Neuroradiology* 37(3), 394–401 (2016)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* 30 (2017)
20. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 109–119. Springer (2021)
21. Xing, D., Papadakis, N.G., Huang, C.L.H., Lee, V.M., Carpenter, T.A., Hall, L.D.: Optimised diffusion-weighting for measurement of apparent diffusion coefficient (adc) in human brain. *Magnetic resonance imaging* 15(7), 771–784 (1997)
22. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters* 15(5), 749–753 (2018)

# STResNet: Covid-19 Detection by ResNet Transfer Learning and Stochastic Pooling



Wei Wang, Shui-Hua Wang, and Yu-Dong Zhang

**Abstract** Since 2019, COVID-19 has been spreading globally with a very rapid rate of transmission, resulting in a large number of confirmed diagnoses and deaths. The main interdiction measure currently in use for COVID-19 is the isolation of the confirmed population. For this reason, an effective and rapid diagnostic approach is particularly important. In this paper, we propose a deep learning framework (STResNet) for diagnosing COVID-19 from chest CT image slices. The proposed framework uses a modified residual network with 50 network layers as a backbone to extract features from chest CT slices and a support vector machine to classify the extracted features. Experiments show that the proposed framework has excellent performance. In the experiment based on a chest CT slices dataset, STResNet achieved accuracy of  $93.81\% \pm 1.02\%$ , MCC of  $87.64\% \pm 2.02\%$ , FMI of  $93.83\% \pm 0.99\%$ , sensitivity of  $94.03\% \pm 1.07\%$ , precision of  $93.64\% \pm 1.54\%$ , F1-score of  $93.83\% \pm 0.99\%$ , and specificity of  $93.59\% \pm 1.67\%$ . These demonstrate the excellent performance of the proposed framework with well balance and stability.

**Keywords** ResNet-50 · Stochastic pooling · Support vector machine

## 1 Introduction

Since December 2019, a disease caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) infection (COVID-19) has been spreading rapidly around the world [1]. The main symptoms of COVID-19 are fever, cough, loss of taste, lung infection and, in severe cases, death [2]. As of October 2022, the disease has killed more than six million people. The main reasons for this are the highly infectious and unstoppable nature of COVID-19 and its rapid mutation. Currently, the main measure of COVID-19 interruption in various countries and regions is the isolation of confirmed patients. However, a proportion of COVID-19 patients are

---

W. Wang · S.-H. Wang · Y.-D. Zhang (✉)  
School of Computing and Mathematical Sciences, University of Leicester, Leicester LE1 7RH,  
UK  
e-mail: [yudongzhang@ieee.org](mailto:yudongzhang@ieee.org)

asymptomatic, which makes targeted diagnosis difficult. Broad coverage diagnosis is also difficult to implement due to the limitations of current diagnostic methods [3].

Currently, the main diagnostic modalities for COVID-19 are Reverse Transcription Polymerase Chain Reaction (RT-PCR) and expert diagnosis based on chest CT images. Among them, RT-PCR is widely used for community diagnosis due to its fast diagnostic speed and low cost. However, RT-PCR has a high false negative rate and tends to miss infected patients. On the other hand, expert diagnosis based on chest CT images is mainly manual, subjective and difficult to implement on a large scale due to the lack of medical experts. In this context, it is important to explore new diagnostic methods.

With the rapid development of computer technology and the availability of computing resources, computer-aided diagnosis system (CAD) is widely used in the diagnosis of many diseases due to their fast and accurate diagnosis [4]. Consequently, medical image-based diagnostic tasks are one of the most active areas of CAD. Therefore, a variety of studies on the COVID-19 CAD system have been proposed by many researchers.

Khan [5] constructed a COVID-19 CAD system based on a multiplexed data enhancement approach to preprocess a CT image slice dataset with a pseudo-Zernike moment derived from the Zernike moment as the feature extracted from the CT image slice and a deep stacked sparse autoencoder as the classifier. Wang [6] proposed a wavelet entropy-based COVID-19 CAD system. The system extracted wavelet entropy from multiple wavelet decomposition results of chest CT images as the extracted features and used the Cat Swarm Optimisation algorithm to train a feedforward neural network to classify the extracted features. Their approach is novel and has potential. Tang, Wang [7] used ensemble learning to integrate multiple deep neural networks to perform chest CT slice-based diagnosis of COVID-19. The proposed model achieved over 90% accuracy in performance. Gafoor, Sampathila [8] used a convolutional neural network with four convolutional layers to perform a binary classification task on chest CT slices to diagnose COVID-19, achieving up to 94% accuracy. Han, Hu [9] used stationary wavelet entropy as the extracted features and extreme learning machine as the classifier to diagnose chest CT slices. Jiang, Brown [10] used a multiple-distance grey-level cooccurrence matrix to extract image features, a feedforward neural network as a classifier, and a genetic algorithm as a training algorithm to construct a COVID-19 CAD system based on chest CT slices.

In this paper, we proposed a deep learning framework (STResNet) for diagnosing COVID-19 from chest CT image slices. STResNet consists of two main components, a modified 50-layer residual network as the backbone to extract features from chest CT slices and a support vector machine as the head to classify the extracted features. In the rest of the paper, Sect. 2 describes the dataset we used for the experiment; Sect. 3 introduces the components of STResNet; Sect. 4 shows and discusses the performance of STResNet in the proposed experiments.

**Table 1** Statistic of the dataset used for the experiment

Class	Num of subjects	Num of images	Subjects' age range
COVID-19	142	320	22–91
HC	142	320	22–76

## 2 Dataset

The experimental of this paper uses a dataset of chest CT slices proposed by Wang, Govindaraj [11]. The dataset contained a total of 640 samples from 282 subjects. The data samples were divided into two classes, the COVID-19 class and the Health Control class (HC). The COVID-19 category contains 320 chest CT slices from 142 COVID-19 patients aged 22–91 years. The HC category contains 320 chest CT slices from 142 healthy subjects aged 21–76. Detailed statistics of the data set are shown in Table 1.

## 3 Methodology

This paper proposes a novel deep learning-based framework for diagnosing COVID-19 from chest CT image slices. This framework replaces all pooling layers in a Residual Network (ResNet-50) with a stochastic pooling layer, which is used as a backbone to extract features from chest CT slices and a support vector machine (SVM) as a head to classify the extracted features. The overall structure of the network is shown in Fig. 1.

### 3.1 ResNet-50

Convolutional neural networks are widely used in image-processing tasks due to their translation invariance and weight-sharing properties. Theoretically, deeper neural networks can perform better in more complex tasks with more parameters and complex network structures. However, in practice, as the depth of the network increases, the performance of neural networks tends to saturate or even decline. The Residual Network (ResNet) alleviates this problem by adding residual learning and a fast gradient channel to the network [12].



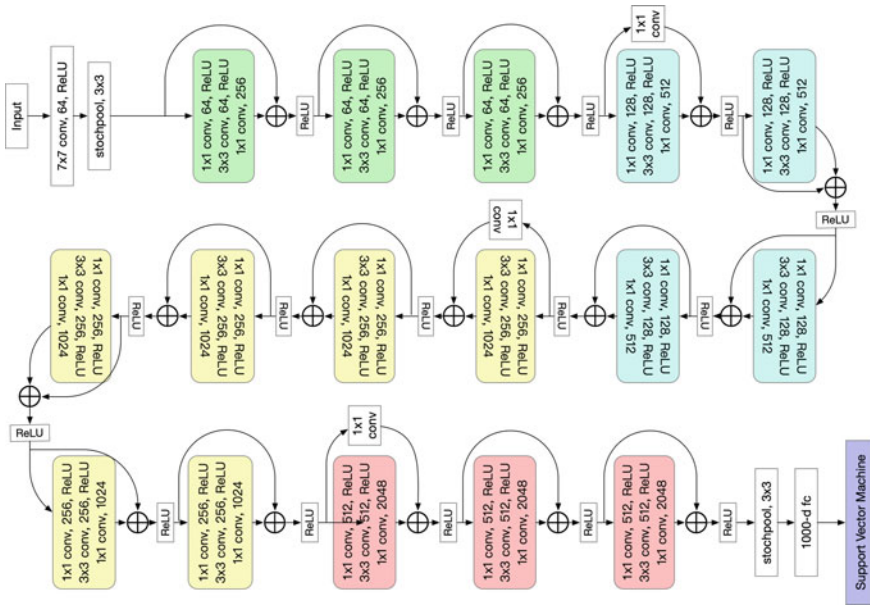
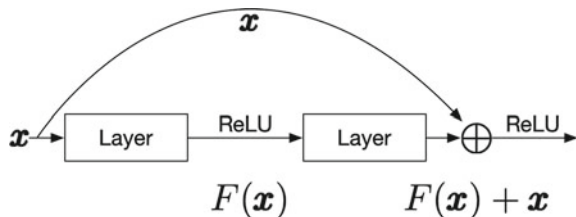


Fig. 1 Overall structure of the proposed framework

ResNet mainly consists of many convolutional blocks, and each block contains several convolutional layers with the rectified linear unit (ReLU) in between. As the name suggests, ResNet is based on residual learning, which solves the problem of degradation of network performance as the network depth increases by having the network learn the residual between the original input and the features learned by the network layer. Therefore, the units of a ResNet are known as residual learning units. The residual learning unit is shown in Fig. 2.

Other than the residual learning units, ResNet contains a  $7 \times 7$  convolutional layer and a  $3 \times 3$  max pooling layer as preprocessing of the input data. In addition, an average pooling layer and a fully connected layer are added to the end of the network to downscale and classify the extracted features.

Fig. 2 Illustration of residual learning unit where  $x$  is the input of the unit, and  $F(x)$  is the features learned by the unit. The output of a residual unit is the output of a ReLU with the combination of  $x$  and  $F(x)$  as input



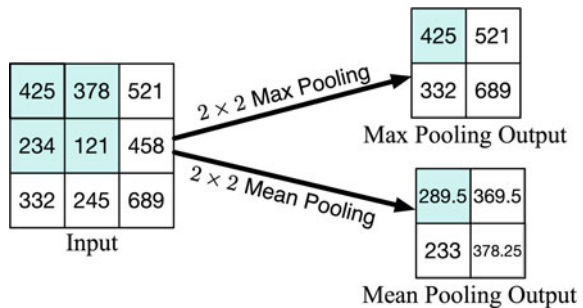
Another possible reason for the saturation of performance in high-depth neural networks is the disappearance of gradients or gradient explosion due to the increased number of layers in the network. As shown in Fig. 1, the fast channel the residual learning units use to transfer their original input allows a direct connection between the different residual learning units in the network. It allows the gradient to be passed directly to each residual learning unit so that the gradient is not degraded by too many layers of the network as the depth of the network increases, leading to degradation or over-enhancement leading to gradient explosion.

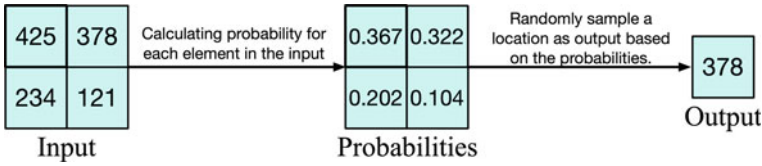
### 3.2 Stochastic Pooling

The pooling layer is an important component of a convolutional neural network that reduces the features' dimensionality while preserving the features' structural information, thus reducing the complexity of the CNN and reducing the time and space cost of the network training. The pooling process is similar to that of the convolutional layer in that the pooling process is performed by sliding a preset size window over the grid data in an orderly manner from left to right and from top to bottom, and pooling is performed during each sliding process. Therefore, pooling layers can be translation invariant, rotation invariant and scale invariant. Different types of pooling layers usually have different pooling operations. The most common pooling layers are the average pooling layer and the maximum pooling layer.

In the average pooling layer, the pooling operation takes the average of the eigenvalues in the area covered by the window during the sliding process, while in the maximum pooling layer, the pooling operation takes the maximum of the eigenvalues in the area covered by the window during the sliding process. Figure 3 shows an example of maximum pooling and average pooling.

**Fig. 3** Illustration of max pooling and mean pooling examples with the same input





**Fig. 4** Illustration of a stochastic pooling operation example

However, in average pooling, it is susceptible to extreme eigenvalues as the average pooling operation considers all eigenvalues. For example, more eigenvalues close to zero will reduce the weight of strong activation. Maximum pooling, which selects only the largest eigenvalues, is a good solution to the above problem of average pooling. However, considering only the largest eigenvalues makes it easy to ignore other valuable eigenvalues, which can lead to overfitting the network. Therefore, the generalization of maximal pooling is generally poor.

The stochastic pool layer [13] is a new type of pooling layer, which solves the problems of traditional pooling by adding randomness to the pooling process to enhance model generalization. The pooling process of the Stochastic pool layer is divided into two steps: (1) calculating the probability of each feature based on the size of the feature in the window and (2) selecting the feature in a weighted random way based on the probability of each feature. Assume that the coverage area  $R_j$  of the pooling window at the  $j$ th slide consists of  $k$  eigenvalues  $(x_1, x_2, \dots, x_k)$ . The computation of the probability of the  $i$ th eigenvalue is shown in Eq. (1).

$$p_i = \frac{x_i}{\sum_{k \in R_j} x_k}. \quad (1)$$

In each sliding process, stochastic pooling selects an eigenvalue based on the probabilities  $P$  of the covered region as the output of the current sliding process, as shown in Eq. (2).

$$A_j = x_l, l \sim P(p_1, \dots, p_{|R_j|}), \quad (2)$$

where  $l$  is the location of the selected feature value. Figure 4 shows an example of stochastic pooling.

### 3.3 Support Vector Machine

The Support Vector Machine (SVM) is a high-performance classification model widely used for various classification tasks [14]. In contrast to most other classification models, SVMs are characterized by the fact that they are not limited to successfully classifying samples but find the hyperplane that can split the samples

of different classes with the maximum interval between each class of samples. Since SVMs include a kernel technique, they are non-linear. Suppose a linearly divisible data set  $T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$  with  $N$  samples. The geometric interval between the hyperplane  $w \cdot \mathbf{x} + b = 0$  and the  $i$ -th sample  $(\mathbf{x}_i, y_i)$  can be calculated by Eq. (3).

$$\gamma_i(w, b) = y_i \left( \frac{w}{\|w\|} \cdot \mathbf{x}_i + \frac{b}{\|w\|} \right). \quad (3)$$

In this case, the distance  $\gamma$  between the hyperplane and the nearest sample is given by Eq. (4).

$$\gamma(w, b) = \min_{i=1,2,\dots,N} \gamma_i(w, b). \quad (4)$$

According to the learning strategy of maximizing the interval of an SVM, the solution of an SVM can be expressed as a constrained optimization problem represented by Eq. (5).

$$\max_{w,b} \gamma(w, b). \quad (5)$$

## 4 Experiment Results and Discussion

### 4.1 Experiment Results of STResNet

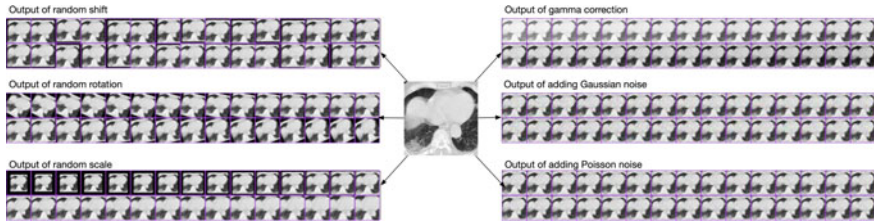
A tenfold cross-validation was introduced in the experimental section to ensure that the data were fully utilized. Specifically, the dataset was randomly divided into 10 data sets with the same amount of data without any return. Since the total data volume could not be divided into ten parts, the tenth data set had slightly less data than the other nine data sets. The experiment consisted of ten runs, with one data set selected as the test set for each run to evaluate the model performance and the other nine data sets as the training set to train the model. The final performance of the model was determined by calculating the mean and standard deviation of the performance obtained over the ten runs. Seven performance metrics were used to evaluate the performance of the model, namely Sensitivity (Sen), Specificity (Sp), Precision (Pre), Accuracy (Acc), F1-score (F1), Matthews Correlation Coefficient (MCC) and the Fowlkes-Mallows Index (FMI). Table 2 shows the experimental results of STResNet in 10 runs and the mean & standard deviation (MSD) of the 10 runs results.

STResNet has a very good and stable overall performance in the COVID-19 diagnostic task based on chest CT slices, achieving an accuracy of  $93.81\% \pm 1.02\%$ ,

**Table 2** Experiment results of STResNet in 10 runs and mean and standard deviation (MSD)

Run	Sen	Spc	Prc	Acc	F1	MCC	FMI
1	92.50	92.81	92.79	92.66	92.64	85.31	92.64
2	93.44	93.75	93.73	93.59	93.58	87.19	93.58
3	92.19	93.75	93.65	92.97	92.91	85.95	92.92
4	94.69	94.69	94.69	94.69	94.69	89.38	94.69
5	94.69	90.00	90.45	92.34	92.52	84.78	92.54
6	94.38	94.69	94.67	94.53	94.52	89.06	94.52
7	93.75	95.62	95.54	94.69	94.64	89.39	94.64
8	94.38	92.19	92.35	93.28	93.35	86.58	93.36
9	94.69	93.12	93.23	93.91	93.95	87.82	93.96
10	95.62	95.31	95.33	95.47	95.48	90.94	95.48
MSD	94.03 ± 1.07	93.59 ± 1.67	93.64 ± 1.54	93.81 ± 1.02	93.83 ± 0.99	87.64 ± 2.02	93.83 ± 0.99

MSD: mean ± standard deviation; Sen: sensitivity; Spc: specificity; Prc: precision; Acc: accuracy; F1: F1-score; MCC: Matthews correlation coefficient; FMI: Fowlkes-Mallows index



**Fig. 5** Examples of data augmentation output

MCC of  $87.64\% \pm 2.02\%$ , and FMI of  $93.83\% \pm 0.99\%$ . The sensitivity of  $94.03\% \pm 1.07\%$ , precision of  $93.64\% \pm 1.54\%$ , and F1-score of  $93.83\% \pm 0.99\%$  obtained by the model show that STResNet has an excellent diagnostic capability for positive samples, i.e., it can accurately diagnose COVID-19 patients from chest CT slices. The model obtained a specificity of  $93.59\% \pm 1.67\%$ , showing that the model can accurately identify healthy subjects from chest CT slices. Overall, the performance of STResNet in the chest CT slice-based COVID-19 diagnosis task was well balanced, with the excellent diagnostic ability for both COVID-19 patients and healthy subjects.

## 4.2 Data Augmentation Results

In order to improve the generalizability of the model and reduce the negative impact of the small collective size of the data, six data augmentation techniques were introduced in this paper, namely shift, rotation, scale, gamma correction, adding Gaussian noise and adding The output of each augmentation is shown in Fig. 5.

## 4.3 Stochastic Pooling Against Max Pooling and Average Pooling

To verify the contribution of stochastic pooling to STResNet performance. We conducted an ablation experiment based on the same dataset, which uses SVM as the head, and the original ResNet-50 and STResNet as the backbone, respectively. The original ResNet-50 contains a max pooling layer at the beginning of the network during the input data preprocessing phase and an average pooling layer before the fully connected layer at the end of the network. In STResNet, both pooling layers are replaced with stochastic pooling layers. Table 3 and Fig. 6 show the model's performance with the original ResNet-50 as the backbone and SVM as the head. Compared with Table 2, it can be seen that STResNet performs better than the model with the original ResNet-50 as the backbone with SVM as the head among all performance metrics.

**Table 3** Experiment results of ResNet-50 with SVM

Run	Sen	Spc	Prc	Acc	F1	MCC	FMI
1	93.75	91.88	92.02	92.81	92.88	85.64	92.88
2	93.12	92.81	92.83	92.97	92.98	85.94	92.98
3	92.19	91.25	91.33	91.72	91.76	83.44	91.76
4	91.25	92.50	92.41	91.88	91.82	83.76	91.83
5	92.81	93.75	93.69	93.28	93.25	86.57	93.25
6	93.44	93.75	93.73	93.59	93.58	87.19	93.58
7	92.19	89.69	89.94	90.94	91.05	81.90	91.06
8	92.81	91.88	91.95	92.34	92.38	84.69	92.38
9	92.81	91.88	91.95	92.34	92.38	84.69	92.38
10	91.25	92.50	92.41	91.88	91.82	83.76	91.83
MSD	92.56 ± 0.84	92.19 ± 1.20	92.23 ± 1.11	92.37 ± 0.81	92.39 ± 0.79	84.76 ± 1.61	92.39 ± 0.79

MSD: mean ± standard deviation; Sen: sensitivity; Spc: specificity; Prc: precision; Acc: accuracy; F1: F1-score; MCC: Matthews correlation coefficient; FMI: Fowlkes-Mallows index

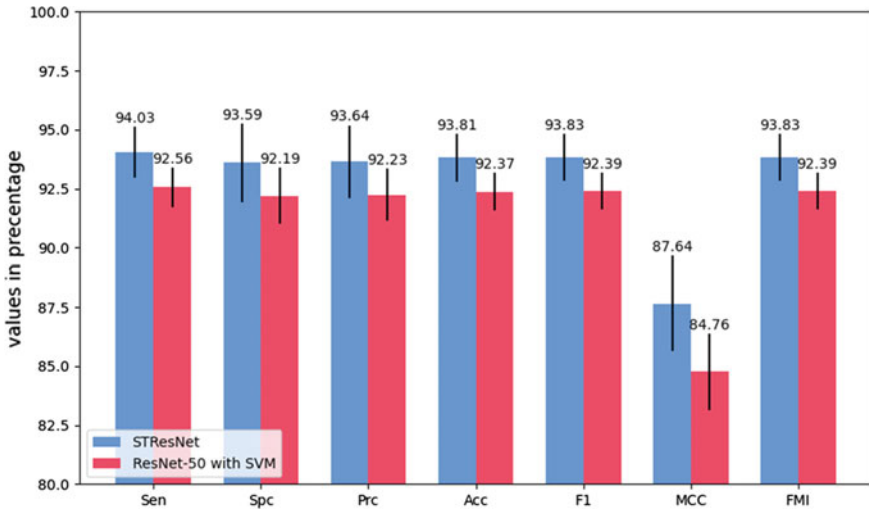


Fig. 6 Performance comparison between STResNet and RestNet-50 with SVM

#### 4.4 Comparison to State-to-the-Art Approaches

Table 4 shows the performance of STResNet compared to other state-of-the-art (SOTA) CAD systems based on chest CT images. STResNet improves in all performance metrics, confirming the relevance of the proposed approach for COVID-19 diagnosis based on chest CT slices.



**Table 4** Comparison to state-of-the-art approaches

Model	Sen	Spc	Prc	Acc	F1	MCC	FMI
PZM-DSSAE [5]	92.06 ± 1.54	92.56 ± 1.06	92.53 ± 1.03	92.31 ± 1.08	92.29 ± 1.10	84.64 ± 2.15	92.29 ± 1.10
WE-CSO [6]	74.06 ± 2.96	78.06 ± 1.81	77.17 ± 1.17	76.06 ± 1.18	75.55 ± 1.58	52.21 ± 2.28	75.58 ± 1.54
EDL-COVID [7]	93.41 ± 0.99	92.81 ± 1.30	92.87 ± 1.22	93.11 ± 0.93	93.13 ± 0.92	86.23 ± 1.86	93.13 ± 0.92
DLM [8]	87.37 ± 1.51	88.12 ± 1.94	88.06 ± 1.75	87.75 ± 1.31	87.71 ± 1.29	75.52 ± 2.62	87.71 ± 1.29
SWE-ELM [9]	82.53 ± 1.23	81.88 ± 1.97	82.03 ± 1.51	82.20 ± 0.85	82.26 ± 0.76	64.43 ± 1.70	82.27 ± 0.76
MDGLCM-GA [10]	81.56 ± 2.25	80.84 ± 1.76	80.99 ± 1.62	81.20 ± 1.65	81.26 ± 1.71	62.42 ± 3.30	81.27 ± 1.71
STResNet (ours)	94.03 ± 1.07	93.59 ± 1.67	93.64 ± 1.54	93.81 ± 1.02	93.83 ± 0.99	87.64 ± 2.02	93.83 ± 0.99

MSD: mean ± standard deviation; Sen: sensitivity; Spc: specificity; Prc: precision; Acc: accuracy; F1: F1-score; MCC: Matthews correlation coefficient; FMI: Fowlkes-Mallows index

## 5 Conclusion

This paper proposes a deep learning framework as a COVID-19 CAD system (STResNet) using a residual learning network with 50 network layers containing stochastic pooling as the backbone and an SVM classifier as the head. The proposed approach has experimentally demonstrated excellent performance in diagnostic tasks based on chest CT slices. STResNet is theoretically applicable to other medical image classification tasks. However, this requires further experimental validation. In the future, we will further experiment with algorithms that can automatically filter different backbones to build a CAD system that can be used for the diagnosis of multiple diseases.

**Acknowledgements** This paper is partially supported by Medical Research Council Confidence in Concept Award, UK (MC PC 17171); Royal Society International Exchanges Cost Share Award, UK (RP202G0230); Hope Foundation for Cancer Research, UK (RM60G0680); Global Challenges Research Fund (GCRF), UK (P202PF11); Sino-UK Industrial Fund, UK (RP202G0289); British Heart Foundation Accelerator Award, UK (AA/18/3/34220); LIAS Pioneering Partnerships award, UK (P202ED10); and Data Science Enhancement Fund, UK (P202RE237).









## References

1. Bchetnia, M., et al., *The outbreak of the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): A review of the current global status*. Journal of Infection and Public Health, 2020. **13**(11): p. 1601–1610.
2. Elibol, E., *Otolaryngological symptoms in COVID-19*. European Archives of Oto-Rhino-Laryngology, 2021. **278**(4): p. 1233–1236.
3. Dasgupta, A., et al., *Epidemiological challenges in pandemic coronavirus disease (COVID-19): Role of artificial intelligence*. WIREs Data Mining and Knowledge Discovery, 2022. **12**(4): p. e1462.
4. Yanase, J. and E. Triantaphyllou, *A systematic survey of computer-aided diagnosis in medicine: Past and present developments*. Expert Systems with Applications, 2019. **138**: p. 112821.
5. Khan, M.A., *Pseudo Zernike Moment and Deep Stacked Sparse Autoencoder for COVID-19 Diagnosis*. CMC-Computers, Materials & Continua, 2021. **69**(3): p. 3145–3162.
6. Wang, W., *Covid-19 Detection by Wavelet Entropy and Cat Swarm Optimization*. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 2022. **415**: p. 479–487.
7. Tang, S.J., et al., *EDL-COVID: Ensemble Deep Learning for COVID-19 Case Detection From Chest X-Ray Images*. IEEE Transactions on Industrial Informatics, 2021. **17**(9): p. 6539–6549.
8. Gafoor, S.A., et al., *Deep learning model for detection of COVID-19 utilizing the chest X-ray images*. Cogent Engineering, 2022. **9**(1).
9. Han, X., et al., *COVID-19 Diagnosis by Stationary Wavelet Entropy and Extreme Learning Machine*. International Journal of Patient-Centered Healthcare, 2022. **12**(1): p. 309952.
10. Jiang, X., et al., *COVID-19 Diagnosis by Multiple-Distance Gray-Level Cooccurrence Matrix and Genetic Algorithm*. International Journal of Patient-Centered Healthcare, 2022. **12**(1): p. 309951.
11. Wang, S.-H., et al., *Covid-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network*. Information Fusion, 2021. **67**: p. 208–229.

12. He, K., et al. *Deep Residual Learning for Image Recognition*. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
13. Zeiler, M.D. and R. Fergus *Stochastic Pooling for Regularization of Deep Convolutional Neural Networks*. 2013. [arXiv:1301.3557](https://arxiv.org/abs/1301.3557).
14. Suthaharan, S., *Support Vector Machine*, in *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*. 2016, Springer US: Boston, MA. p. 207–235.

# Convolutional Neural Networks for Newborn Pain Assessment Using Face Images: A Quantitative and Qualitative Comparison



Gabriel A. S. Coutrin , Lucas P. Carlini , Leonardo A. Ferreira , Tatianny M. Heiderich , Rita C. X. Balda , Marina C. M. Barros , Ruth Guinsburg , and Carlos E. Thomaz 

**Abstract** Pain experience, when intense or repetitive, may harm the development of newborns. Several clinical and non-clinical studies have been carried out to identify the presence of pain through behavioural analysis, mainly by facial mimicry. Advances in deep learning might show automatic, continuous and non-invasive solutions for neonatal pain assessment as well. In this context, this work investigates the following five state-of-the-art Convolutional Neural Networks (CNNs) for the classification of pain using two distinct face image datasets (UNIFESP and iCOPE): VGG-16, ResNet50, SENet50, and Inception-V3, all implemented with transfer learning, and the specific one called Neonatal CNN, which was trained end-to-end. Our experimental results, based on quantitative and qualitative analyses, indicate the superiority of models originally trained with face images, highlighting most relevant differences owing to the explainable information extracted by each model and the current issue of limited neonatal face images available.

**Keywords** Neonatal pain · Facial expression · Deep learning

## 1 Introduction

The International Association for the Study of Pain describes pain as an “unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage” [1]. The subject that experiences pain tends to perform an act of escape and withdrawal from the source of this phenomenon [2]. If not avoided, the constant presence of pain increases the suffering of a certain

---

G. A. S. Coutrin (✉) · L. P. Carlini · L. A. Ferreira · T. M. Heiderich · C. E. Thomaz  
University Center of FEI, São Bernardo do Campo, São Paulo, SP, Brazil  
e-mail: [gcoutrin@outlook.com](mailto:gcoutrin@outlook.com)

C. E. Thomaz  
e-mail: [cet@fei.edu.br](mailto:cet@fei.edu.br)

R. C. X. Balda · M. C. M. Barros · R. Guinsburg  
Federal University of São Paulo, São Paulo, SP, Brazil

individual and also results in the reduction of their life span [3]. Consequently, the assessment and correct treatment are mandatory for the healthy development of human beings.

Specifically about neonatal pain, due to the inability of newborn babies to indicate pain by verbal communication, it was believed in the past that the central nervous system of neonates was not fully developed, and, as a consequence, painful procedures were carried out with insufficient (or none) analgesics for the relief of pain [4]. Fortunately, since 1980s, several studies have shown that, during gestation, the developing nociceptive system is able to process painful stimulus [4–6]. Moreover, it was reported that the pain experienced by critically ill neonates is associated with changes in their cardiovascular, respiratory and metabolic stability, which increases mortality in neonatal intensive care units [7]. Therefore, these studies demonstrate that reliable and precise pain assessment tools are mandatory for their healthy development and well-being.

Since then, several clinical scales have been developed. Recently, Tamanaka et al. [8] carried out a systematic literature review that identified 52 scales published from 1971 to 2020 that are based on the facial expression responses for neonatal pain assessment. This work also reported that the eyes, the region in-between eyebrow, forehead, nasolabial furrows, and mouth are the facial features most analysed by clinical scales. Meanwhile, computer scientists have proposed several automatic frameworks, also based on facial expression, which enable continuous monitoring of the newborn and are specific to the pain phenomena [9–14]. Mainly, these frameworks applied transfer learning on Convolutional Neural Networks (CNNs) that were originally trained on facial or object recognition. Specifically, Zamzmi et al. [14] proposed the Neonatal Convolutional Neural Network (N-CNN), the first CNN architecture that was implemented end-to-end to neonatal pain assessment. Using a private dataset [14] and the public iCOPE one [9], the authors compared the N-CNN with the ResNet architecture and with a classifier based on Local Binary Patterns (LBPs) and SVMs. Regarding the first dataset, the N-CNN achieved 91% accuracy and 0.93 AUC, while the ResNet reached 87.1% accuracy and 0.89 AUC, and the LBP obtained 85.5% accuracy and 0.82 AUC. To the iCOPE, the N-CNN, ResNet and LBP achieved, respectively, 84.5%, 82.9%, and 81.3% accuracy.

Even though these frameworks implemented several CNNs architectures with high classification performance, each work carried out their own training/test procedure using distinct face image datasets. Therefore, preventing a direct comparison of performance between these architectures. In this context, the current work presents two main contributions: (I) we propose a novel training/test protocol named *leave-some-subject-out*, based on the original *leave-one-subject-out*, where we create folds splitting training-test sets according to the identity of the neonate; and (II) we implement and compare a number of state-of-the-art (SOTA) CNNs, evaluating their performance based not only on quantitative but also qualitative analyses using eXplainable Artificial Intelligence (XAI).

The remainder of this work can be summarised as follows. In Sect. 2, we describe the face image datasets used, classification models, the proposed training protocol, and the XAI technique implemented. Next, in Sect. 3, we show our quantitative and

qualitative experimental results, and, then, we discuss the performance of each classification model. Finally, in Sect. 4, we conclude our work, discuss the impact of our findings, and provide guidance for future research.

## 2 Materials and Methods

In this section, we firstly describe the face image datasets used, labelled as neonates with “pain” or “no pain”. Next, we depict the classification models that are being investigated, and their training/test protocols. Then, we describe the Gradient-weighted Class Activation Mapping (Grad-CAM) [15], that is, the XAI algorithm implemented here to qualitatively analyse the classification results.

### 2.1 Face Images Datasets

**UNIFESP Image Dataset:** Heiderich et al. [12] developed the UNIFESP dataset at the Federal University of São Paulo. It contains 360 face images with resolution of  $320 \times 233$ , which were captured from 30 neonates with 24 to 168 h of life. These photographs were taken during routine painful procedures, such as intramuscular or capillary injection and venipuncture. Each image was randomly evaluated by a group of health professionals with experience in neonatal intensive care units. The assessment resulted in 164 “pain” images and 196 “no pain” ones.

**iCOPE Dataset:** Brahnma et al. [9] created the infant Classification of Pain Expression (iCOPE) dataset. A total of 200 face images, with resolution of  $3008 \times 2000$ , were captured from 26 neonates with 18 h to 3 days of life, all Caucasians. The photographs were taken during a session in which the neonates experienced 4 different stimuli, in the presented order: transport between cribs, air stimulus, friction on the heel using a cotton wool; heel puncture for blood collection (painful stimuli). Thus, the iCOPE images are divided in: 18 images of neonates crying, 36 of heel friction, 23 of air stimulation, 63 resting neonates images and 60 examples of neonates during a painful procedure. For the present work, only the images classified as “pain” and “rest” were used.

### 2.2 Classification Models

For the classification task, we employed the following SOTA CNN architectures that were selected based upon their accuracy on pre-trained datasets for facial recognition or ImageNet if the model didn't have any available pre-training on face images:

**Table 1** Fully connected layers used for each pre-trained model.

Model	VGG-16	ResNet50	SENet50	Inception-V3
FC 1	512, ReLU	1000, ReLU	1000, ReLU	512, ReLU
FC 2	512, ReLU	–	–	512, ReLU
Output	2, Softmax	2, Softmax	2, Softmax	2, Softmax
Total parameters	27,823,938	25,612,154	28,143,146	23,115,554
Fine tuning conv. layers	6	9	9	18

- **VGG-16:** Parkhi, Vedaldi, and Zisserman [16] implemented a VGG-16 architecture [17] for facial recognition. The CNN was trained on 2.6 millions face images from 2622 individuals (VGGFace dataset)<sup>1</sup>;
- **ResNet50:** Cao et al. [19] extended the original VGGFace database, originating a new dataset containing 3.3 millions face images. The ResNet50 [20] model trained on it and is capable of recognising 9131 individuals<sup>1</sup>;
- **SENet50:** Following the ResNet50, the “Squeeze-and-Excitation” ResNet50 architecture [21] was also trained on the VGGFace2 dataset [19]<sup>1</sup>;
- **Inception-V3:** For the Inception-V3 architecture [22], no pre-trained model for face recognition was found. Therefore, we used the model trained on the ImageNet dataset;
- **N-CNN:** In addition to the aforementioned pre-trained models, the N-CNN of Zamzmi et al. [14] was also implemented. Since no pre-trained version of the model was found, this work rebuilt the N-CNN (as described in [14]) and carried out its training with the iCOPE and UNIFESP datasets.

As mentioned in the previous Sect. 2.1, due to the limited sample size of both UNIFESP and iCOPE datasets, we applied the well-known transfer learning strategy. Therefore, the pre-trained models were selected, refined and submitted to a new task of pain classification in newborns. For each pre-trained CNN, the original convolutional layers were preserved and a new set of fully connected (FC) layers were attached to the top of the CNN. In order to maximise its accuracy, the configuration of the new sequence of FC layers was defined experimentally, with parameters ranging from 50 to 2048 neurons and from 1 to 3 layers. Also, we fine-tuned the last convolutional layers, where their weights were updated during the new training with newborn images. Table 1 shows the adapted models.

---

<sup>1</sup> We used the pre-trained models available at [18].

### 2.3 Training/Test Protocol

Before training each model, we extracted the face of each image. The cropped images intended for training the models were submitted to a *data augmentation* process, generating 20 new samples per image by randomly applying the following manipulations: rotation (30°), shear (0.15), width and height shift (0.20), zoom (0.70–1.5), brightness (0.50–1.1) and horizontal flip. Also, for each newly generated image, we verified whether that face was still automatically detectable (that is, visible and within the image bounds).

All models were trained using the RMSprop optimizer [23] and with Categorical Cross-Entropy loss. To speed up training and reduce the possibility of models being overfitted, we used the mini-batch strategy, dividing the data into batches of 16 images. For the new FC layers, we applied the Dropout (0.5) and the weight regularisation  $l_1$  ( $5 \times 10^{-4}$ ). The learning rate  $\eta$  was dynamically altered from  $\eta = 1 \times 10^{-4}$  during training to  $\eta = 1 \times 10^{-4}$  during fine-tuning.

Initially, training was performed only to adjust the weights of the new set of fully connected layers, except for N-CNN, in which all layers needed training. For the pre-trained models, if there is no error reduction for 5 consecutive epochs (processing of all available images), fine-tuning of the convolutional layers starts. Therefore, the weights of the last convolutional layers are updated as well (Table 1). For models in the fine-tuning phase and for N-CNN, training was terminated after 10 consecutive epochs without loss reduction.

All CNNs were trained and tested with the union of UNIFESP and iCOPE databases, using the *leave-some-subjects-out* protocol: the dataset subjects are equally divided into folds and, at each iteration of the cross-validation, one fold is preserved for testing the model, whereas the others are used in the training. In this way, there are more images for testing (when compared to the traditional *leave-one-subject-out*) and information leakage is avoided, since there will not be images of the same subject in both training and test sets at the same time. Then, for the total of 56 subjects (30 from UNIFESP and 26 from iCOPE), 10 folds were created, each containing 5 or 6 subjects (3 necessarily from UNIFESP).

### 2.4 Explainable Artificial Intelligence

XAI techniques are able to explain the decision-making process that leads an AI model to a particular answer, allowing a sound understanding of such models. Recently, Velden et al. [24] presented a systemic review of works that used deep learning-based XAI in medical image analysis, categorising their explanation method into three types: visual, textual, and example-based. As shown by the authors, the most popular XAI method is the Grad-CAM [15].

Grad-CAM evaluates the gradients between the final convolution layer and the desired output. In fact, Grad-CAM is able to produce an attribution mask in terms



of any convolutional layer of the CNN. However, as stated by Selvaraju et al. [15], it is expected that the last convolutional layers have the best compromise between detailed spatial information and high-level semantics, since its neurons look for semantic class-specific information in the image.

To obtain the attribution mask  $L_{\text{Grad-CAM}}^c$  to a specific  $c$  class of a given input  $x$ , it is firstly needed to compute the gradient  $\frac{\partial y^c}{\partial A_{ij}^k}$  of the score  $y_c$  (before the Softmax) with respect to feature maps  $A$  of the desired convolutional layer  $k$ . Then, the importance weights  $\alpha_k^c$  of each feature map are calculated by the global average pooling of the gradients. The final attribution mask is obtained by Eq. 1. ReLU is applied to filter out feature maps that show a negative influence on the class of study. These features are likely to influence other classes instead of the desired one.

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left( \underbrace{\sum_k \alpha_k^c A^k}_{\text{linear combination}} \right), \text{ where} \tag{1}$$

$$\alpha_k^c = \underbrace{\frac{1}{Z} \sum_i \sum_j}_{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backpropagation}}$$

### 3 Results and Discussion

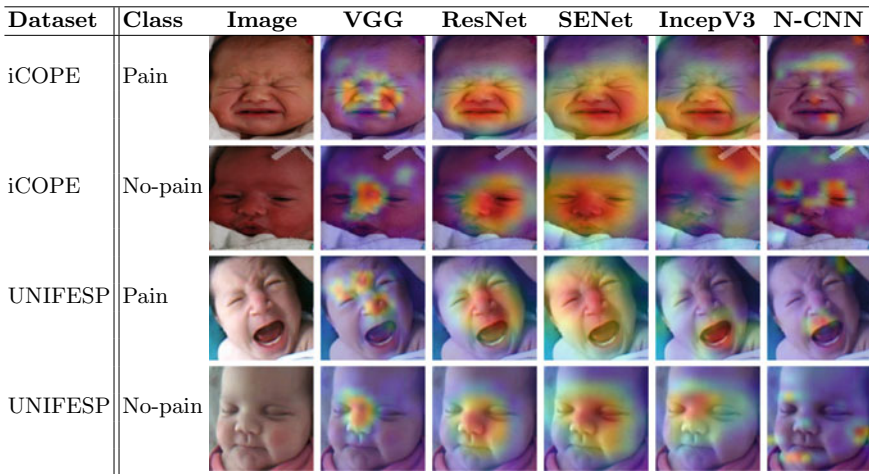
In this section, we begin analysing the quantitative results of each classification model regarding their evaluation metrics. Then, we show the qualitative results provided by a XAI technique. Finally, we close discussing our findings.

#### 3.1 Quantitative Results

Table 2 shows the average quantitative performance of each classification model. We can see that the VGG-16, ResNet50, and SENet50 models presented similar performance and outperformed the Inception-V3 and N-CNN ones. The former three CNNs achieved more than 85% level for all the evaluation metrics. There is though no statistical difference between the results of VGG-16, ResNet50, and SENet50 (considering  $p = 0.05$ ). However, VGG-16 has overall the fastest convergence training process: 34 epochs, whereas ResNet50 and SENet50 44 and 40 epochs, on average, respectively. Also, all the CNN models presented a high standard deviation for all the metrics considered ( $\geq 5\%$ ), showing sensitiveness to the training/test samples chosen.

**Table 2** Evaluation metrics results for each model.

Metric	VGG-16	ResNet50	SENet50	Inception-V3	N-CNN
Accuracy	86.2% ± 7%	85.6% ± 7%	86.1% ± 5%	81.3% ± 5%	77.1% ± 7%
F1	87.7% ± 6%	87.0% ± 7%	87.3% ± 5%	84.1% ± 4%	80.8% ± 6%
AUC	85.4% ± 8%	85.2% ± 7%	85.7% ± 5%	80.6% ± 6%	76.0% ± 7%
Epochs	34 ± 7	44 ± 12	40 ± 10	47 ± 12	33 ± 17



**Fig. 1** Qualitative examples of Grad-CAM results for each model.

### 3.2 Qualitative Results

In order to apply the Grad-CAM, for each architecture, we selected the layer before the last pooling layer to generate the class activation maps. Figure 1 shows the results for each CNN using as examples the correct classification of 4 neonate images.

All the models learned to detect discriminant facial regions to pain assessment. Apparently, VGG-16 associates the “no pain” state only with the nose and the region between eyebrows. When “pain” is detected, the model highlights, in addition to the already mentioned regions, the nasolabial furrows and the forehead. Both ResNet50 and SENet50 seem to use the entire face in the decision process. However, it is not clear what is the relation between regions and classes. Inception-V3 is capable of recognising face elements, but its attention is drawn by artefacts, such as the ornament in the second image (from top to bottom). Also, this model seems to associate the pain state with the neonate’s open mouth. Lastly, N-CNN appears to associate the “pain” state with open mouth and the “no pain” state with open eyes. However, its activation maps have a scattered behaviour, which intensifies when these indicators are not present, making it difficult to identify any pattern of facial expression analysis.

### 3.3 Discussion

The quantitative results demonstrated the superiority of the VGG-16, ResNet50 and SENet50 models. We hypothesize that this superiority can be attributed to the original training of each CNN: the three superior models were trained on face images, that is, a datatype related to the iCOPE and UNIFESP datasets. Statistically, there is no significant difference between the quantitative metrics of these three CNNs. However, the qualitative results distinguish the behaviour developed by each CNN for the classification of facial expressions. VGG-16's strategy is focused on the analysis of the nose, the nasolabial furrows, the forehead, and the region between eyebrows of the neonate. These facial features are commonly used by neonatologists [25]. ResNet50 and SENet50, on the other hand, do not focus on specific face regions, but rather use the entire face for classification. Despite achieving similar metrics to the VGG-16, the activation maps generated by this holistic approach make it difficult to interpret what these networks consider "pain" or "no pain". Thus, under the conditions of the present investigation, we consider VGG-16 to be the best suited model for neonatal pain assessment, since it combines high performance with better explainability.

Inception-V3 was pre-trained with the ImageNet dataset, which is made of images from 1000 object categories. Thus, possibly, the convolutions applied to the neonate images did not extract specific features for the classification of facial expressions. This hypothesis is reinforced by the Grad-CAM results. Although Inception-V3 have shown to be capable of using facial elements, the presence of artefacts in the image receives most of the CNN's attention, such as the ornament of the example depicted in the previous section (Fig. 1). This model correctly classified the image, but its decision was not based on the neonate's face.

The implemented N-CNN model had no prior training. Considering that deep learning methods require a large amount of data, the union of the UNIFESP and iCOPE dataset was certainly not enough for training the N-CNN's layers. In the original publication of this architecture [14], the CNN was trained with 3026 images, plus a process of data augmentation, to achieve an average accuracy of 91%. In the qualitative analysis, for some samples, the N-CNN showed exclusive attention to the mouth and eyebrows of the neonate. However, in general, its activation maps are scattered, and no well-defined strategy for image classification is evidenced. In fact, in some cases, the model highlighted regions outside the face of the neonate, such as hair, the contour of the face, and other elements of the image. These observations are consistent with the quantitative results, and reinforce the hypothesis that the N-CNN was not trained with enough images, since it did not learn a clear strategy for recognizing pain patterns.

The lack of images is actually a limiting factor for the application of deep learning in neonatal pain assessment [10, 11]. Besides data protection policies, the very nature of the studied task represents an obstacle for data collection, since painful procedures implies ethical issues. Another aggravating point is the lack of standardization of datasets. This work has united two datasets, but it is important to highlight a possible

incompatibility between the data, for presenting different resolutions and for being labelled according to different methods.

## 4 Conclusion

This paper presented a systematic and detailed comparison of several CNNs when performing neonatal pain assessment. To date, several works have proposed distinct models that achieved SOTA performance, however each framework was evaluated using distinct face image datasets and training/test protocols, consequently, preventing a direct comparison of performance between these architectures.

Firstly, we introduced the *leave-some-subjects-out* training protocol. This strategy avoids data leakage from the training set to the test one, and also better evaluation of the generalisation capabilities of each classification model. Then, we quantitatively and qualitatively compared these models. The quantitative results showed superiority of the VGG-16, ResNet50, and SENet50 over the Inception-V3 and N-CNN models. However, when applying the Grad-CAM XAI technique, we observed that the ResNet50 and SENet50's feature extraction is based on the entire face of the neonate, without distinctions for "pain" and "no pain" images. On the other hand, the VGG-16 model seems to focus on the nasolabial furrow and on the forehead when assessing "pain" images. These regions are indeed clinically relevant and agree with the visual perception of adults when assessing pain [25]. Therefore, we consider the VGG-16 as the best model to combining high classification performance with better explainability.

Finally, as future work, we intend to implement a CNN architecture that, during training, also consider the human knowledge. To accomplish this, we will use the visual cognitive perception of experts performing neonatal pain assessment through facial expression. The combination of prior knowledge of experts with the current SOTA CNN models may enable even better performance and specificity to the pain assessment, overcoming the limited training sample problem and the subjectivity of human judgement as well.

**Acknowledgements** The authors would like to thank the financial support provided by the Brazilian funding agencies CNPq (401059/2019-7), FAPESP (2018/13076-9) and CAPES.

## References

1. IASP. "IASP Publication, Pain terms: a list with definitions and notes on usage". In: Pain (1979).
2. Luda Diatchenko et al. "Genetic architecture of human pain perception". In: TRENDS in Genetics 23.12 (2007), pp. 605–613.
3. Luda Diatchenko et al. "Idiopathic pain disorders-pathways of vulnerability". In: Pain 123.3 (2006), pp. 226–230.

4. Kanwaljeet JS Anand, Paul R Hickey, et al. "Pain and its effects in the human neonate and fetus". In: *N Engl J Med* 317.21 (1987), pp. 1321–1329.
5. Kanwaljeet JS Anand and David B Carr. "The neuroanatomy, neurophysiology, and neurochemistry of pain, stress, and analgesia in newborns and children". In: *Pediatric Clinics of North America* 36.4 (1989), pp. 795–822.
6. Ruth VE Grunau and Kenneth D Craig. "Pain expression in neonates: facial action and cry". In: *Pain* 28.3 (1987), pp. 395–410.
7. Ruth Guinsburg. "Avaliação e tratamento da dor no recém-nascido". In: *J Pediatr (Rio J)* 75.3 (1999), pp. 149–60.
8. Fernanda G. Tamanaka et al. "Neonatal pain assessment: A Kendall analysis between clinical and visually perceived facial features". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization 0.0* (2022), pp. 1–10. <https://doi.org/10.1080/21681163.2022.2044909>.
9. Sheryl Brahnman et al. "Machine recognition and representation of neonatal facial displays of acute pain". In: *Artificial intelligence in medicine* 36.3 (2006), pp. 211–222.
10. Lucas F. Buzuti et al. "Neonatal pain assessment from facial expression using Deep Neural Networks". In: *Anais do XVI Workshop de Visão Computacional (2020)*, pp. 87–92.
11. Lucas P. Carlini et al. "A Convolutional Neural Network-based Mobile Application to Bedside Neonatal Pain Assessment". In: *2021 34th SIB-GRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 2021, pp. 394–401. <https://doi.org/10.1109/SIBGRAPI54419.2021.00060>.
12. Tatiany Marcondes Heiderich, Ana Teresa Figueiredo Stochero Leslie, and Ruth Guinsburg. "Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements". In: *Acta Paediatrica* 104.2 (2015), e63–e69.
13. Ghada Zamzmi et al. "Neonatal pain expression recognition using transfer learning". In: *arXiv preprint arXiv:1807.01631* (2018).
14. Ghada Zamzmi et al. "Pain assessment from facial expression: Neonatal convolutional neural network (N-CNN)". In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–7.
15. Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.
16. Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. "Deep Face Recognition". In: *British Machine Vision Conference*. 2015.
17. Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).
18. Refik Can Malli. *keras-vggface: VGGFace implementation with Keras Frame-work*. [Online; accessed 20-March-2022]. 2016. <https://github.com/rcmalli/keras-vggface>
19. Qiong Cao et al. "Vggface2: A dataset for recognising faces across pose and age". In: *2018 13th IEEE international conference on automatic face and gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.
20. Kaiying He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
21. Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.
22. Christian Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.
23. Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Overview of minibatch gradient descent. [Online; accessed 23-July-2020]. 2012. <http://www.cs.toronto.edu/~tijmen/csc321/slides/lectureslideslec6.pdf>.
24. Bas HM van der Velden et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis". In: *arXiv preprint arXiv:2107.10912* (2021).

25. Lucas Pereira Carlini et al. “A Visual Perception Framework to Analyse Neonatal Pain in Face Images”. In: *Image Analysis and Recognition. Proceedings of the 17th International Conference on Image Analysis and Recognition, ICIAR 2020*. Ed. by Aurélio Campilho, Fakhri Karray, and Zhou Wang. Vol. 12131. *Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2020, pp. 233–243. ISBN: 978-3-030-50347-5.

# Machine Learning for the Evaluation and Detection of Key Markers in Dilated Cardiomyopathy



Xiaodan Bi, Zhenrun Zhan, Jinpeng Yang, Xu Tang, and Tingting Zhao

**Abstract** *Objective:* Screening for dilated cardiomyopathy (DCM) core genes and letter immune infiltration by bioinformatic methods to find new strategies for prevention and treatment. *Methods:* Gene expression compilation database (GEO) GSE3585 and GSE17800 gene microarray sets were extracted and differentially expressed genes (DEGs) were obtained from DCM and normal control myocardial biopsies using R language. The DEGs were tested for gene ontology (GO) functional analysis, Kyoto Gene and Genome Encyclopedia (KEGG) pathway analysis and gene probe (GSEA) enrichment. The Lasso algorithm was subsequently used to identify key DCM-related genes in the practice set and to authenticate them against the test set. Prospective mechanisms for DCM development included: differences in key gene expression between normal and DCM samples, variations in western blot coverage, correlates of clinical relevance to exempt cells, and key gene and immune cell correlation. *Results:* The final screening identified 2 key genes, NPPA and NPPB. The variation in expression of the key genes between normal and DCM samples can be regarded as a diagnostic factor for patients. Also, There is a striking divergence in vaccine levels of immuinia among normal and DCM samples, and the critical gene expression was closely related to the richness of immune cell infiltration. *Conclusion:* Based on bioinformatics analysis and review of relevant literature, two candidate genes, NPPA and NPPB, were screened and strongly associated with the progression of DCM, providing meaningful clues and suggestions used to prevent and heal DCM.

**Keywords** Enrichment analysis · Dilated cardiomyopathy · Bioinformatics · Machine learning · Immuno-infiltration · Differentially expressed genes

---

X. Bi · Z. Zhan · J. Yang · X. Tang · T. Zhao (✉)  
Changzhi Medical College, Changzhi, Shanxi, China  
e-mail: [649823325@qq.com](mailto:649823325@qq.com)

Heping Hospital Affiliated to Changzhi Medical College, Changzhi, Shanxi, China

## 1 Introduction

Dilated cardiomyopathy (DCM) is a cardiomyopathy in which one or both left ventricles have systolic or diastolic dysfunction, in addition to coronary artery disease and abnormal load. The pathogenesis of cardiomyopathy is complex, involving viral infection, immunity, genetics and the environment, and has not been clearly understood [1]. The pathogenesis of cardiomyopathies is complex and not well understood. The pathogenesis of cardiomyopathies is complex, involving viral infections, immunity, genetics and the environment, and has not been clearly understood. So far, Li Guo et al. have identified several related genes encoding myosin, cytoskeleton, nuclear membrane, myosin, Ionic channels and cell to cell connections molecules involved in the pathogenesis of DCM [2]. Genetic mutations encoding myosin (TTN) are thought to be the most common cause of DCM. In addition, transformation of the LMNA gene (nuclear fibre board layer), FLNC (filament protein C), Des (knot protein), PLN (phosphoprotein), and the SCN5A allele has been shown to be a malignant cause of DCM. In this study, comprehensive bioinformatics analysis of endocardial tissue and mRNA export profiles based on the Gene Expression Public Database was conducted to investigate the pathogenesis of the disease, thus providing a reference for molecular mechanism research and clinical diagnosis and treatment exploration.

## 2 Methods and Materials

### 2.1 Data Acquisition and Download

The GPL96-based GeneChip dataset GSE3585 was downloaded to the GEO database (<https://www.ncbi.nlm.nih.gov/>) [3]. GSE358 gene microarray dataset derived among individuals with dilated cardiomyopathy with lowered left septal ejection fraction and normal cardiac function and the GSE17800 gene microarray dataset from the GPL570 platform were downloaded from the GEO database, of which the GSE17800 dataset includes 40 DCM biopsies before IA/IgG treatment following immunosorbency and 8 biopsies for control. The GSE3585 microarray dataset includes seven independent specimens of subendocardial left ventricular tissue from patients with DCM at the moment of transplantation and five NF donor hearts that were not transplanted due to palpable coronary artery calcification due to palpable coronary artery calcification, excluding technical or biological duplicates; for more informative results, only myocardial samples from DCM patients without IA/IgG treatment and myocardial samples from normal hearts were included as criteria for this study Forty specimens from DCM patients prior to IA/IgG treatment and eight specimens from normal hearts from the GSE17800 dataset were therefore obtained.



## ***2.2 Data Processing and Genetic Screening***

Principal component analysis (PCA) was performed on specimens from the above two datasets separately according to different chips using R language to observe the distribution between groups [4, 5]. And used the GEO2R online tool (<https://www.ncbi.nlm.nih.gov/geo/geo2r/>) to profile the aforementioned differentially represented gene in each of the above data sets, and set the screening condition as  $\log_2FC > 1$  or  $\log_2FC < -1$ ,  $P < 0.05$  to identify differentially expressed genes. In contrast, the co-expressed differential genes from both microarrays will contain genes with inconsistent up- and down-regulation, and direct raw signal analysis of all co-expressed differential genes for genes and pathways associated with dilated cardiomyopathy with combined heart failure will confound the effect of false-positive co-expressed genes [4, 6]. To exclude this confounding factor, and in order to screen for genes that can be used as predictive targets for clinical diagnosis and prognosis, it is clear that up-regulated expression genes are more feasible for clinical application and more valuable for study compared to healthy individuals with normal cardiac function, so we selected only those genes that were up-regulated in co-expression differential genes for analysis. The differentially expressed genes obtained from the two datasets were plotted using R language for heat map and volcano map respectively. We also used Venn Diagram to differentially expressed genes were screened obtained from the two datasets for intersection. The differentially up-regulated genes obtained from the two datasets were intersected to obtain differentially up-regulated genes with consistent expression associated with dilated cardiomyopathy.

## ***2.3 Enrichment Analysis: GO, KEGG, DO and GSEA***

The above commonly up-regulated DEGs were extracted, and GO functional analysis and KEGG pathway analysis were performed in R language [7], setting  $P < 0.1$  and adj.  $P < 0.2$  as screening conditions to screen for major enrichment functions and pathways of differential genes [8–10]. The main enabling features and pathways were visualised accordingly, and GSEA enrichment analysis was performed on each of the two chips to explore the main signalling pathways of DCM. The main signalling pathways of the DCM were investigated.

## ***2.4 Selecting and Identifying Gene Predictor Models for Premature Diagnosis***

The data from GSE3585 was chosen as the training set and data from GSE17800 as the test set. LASSO model for GSE3585 was constructed using the glmnet package. Plot the differential gene characteristic curve and calculate its area under the curve

[10, 11]. SVM is a generalised linear classifier that classifies data binary in a supervised learning manner so as to maximise the magnitude of differences using e1071, kernlab and insert symbolic packets to eradicate recurrent traits Data computation on the obtained differential gene to finally obtain the best gene markers [12, 13]. Second machine A learning algorithm was using simultaneously filtering for pivotal kIRC genes to obtain the identical key genes by using R's Venn package. Key genes in GSE3585 were then differentially analysed using the limma package R. In addition, subject operating characteristic (ROC) curves were built and AUC values were generated to assess the predictive value of the results in the exercise and test sets [14].

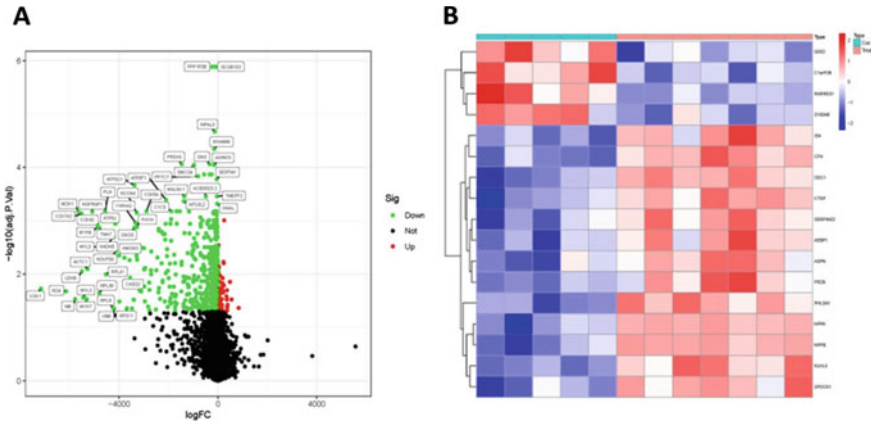
## 2.5 Immune Cell Infiltration Analysis

In this case, immune cells were correlated and a matrix of immune cell infiltration was obtained by Cibersort analysis to assess the percentage of such cells in dilated cardiomyopathy versus normal samples [5]. A heat map was also produced using the ggplot2 package to shed light on the signature of immune cell infiltration in myocardial tissue during heart failure [5, 6, 15, 16]. We also analysed the linkage between the key genes screened and immune cells.

## 3 Results

### 3.1 Screening for DCM-Associated Differential Genes

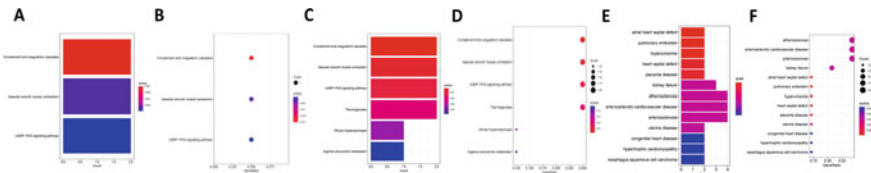
After screening, we obtained 7 independent subendocardial left ventricular tissue specimens from the GSE3580 dataset from DCM patients receiving grafts and 5 NF donation hearts that had not been grafts due to accessible coronary artery calcification, and 40 specimens from the GSE17800 dataset from DCM patients prior to IA/IgG treatment and 8 specimens from normal hearts. The two datasets were subjected to principal component analysis in R and the differential genes were screened for and highly consistent clustering results were obtained for both groups. Based on the screening criteria, the meaningful variably expressed genes were presented in different colours in a volcano plot (Fig. 1a). These results include differential genes that we focused on that were up-regulated in the dilated cardiomyopathy group, and GSE3585 yielded 13 differentially expressed genes that were up-regulated (Fig. 1b).



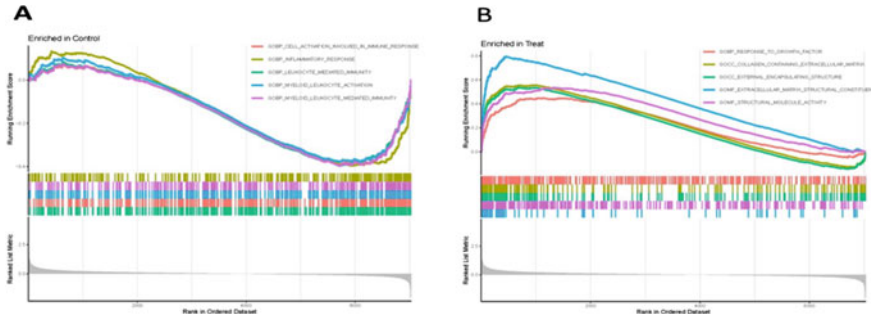
**Fig. 1** **a** The DEG volcano plot, where the up-regulated genes are in red and the down-regulated genes are in green. **b** is a heat map of whole gene expression for normal and DCM samples, where red is high expression and blue is low expression

### 3.2 Co-expressed Genes GO, DO and KEGG Signalling Path Enrichment Analysis

The GO enrichment analysis of the 17 common DEGs obtained from the screening showed that there were 2 cellular components (CC), 81 biological processes (BP), and 29 molecular functions (MF). The results showed that the differential proteins were mainly involved in cell growth regulation, cell growth, astrocyte differentiation, receptor guanylate cyclase signalling pathway and cGMP metabolic process in terms of biological processes; cell composition was mainly distributed in the extracellular matrix containing collagen and neuromuscular junction; molecular function was mainly related to endopeptidase activity (Fig. 2a, b). KEGG enrichment analysis of the above-mentioned differential genes yielded a total of six pathways, the results of which showed that DCM was mainly associated with the complement and coagulation cascades, vascular smooth muscle contraction, and cGMP-PKG signalling pathways (Fig. 2c, d). A total of 13 pathways were obtained by DO analysis. (Fig. 2e, f).



**Fig. 2** Plots **a** and **b** of GO enrichment analysis of 30 genes screened by DEG. plots **c** and **d** of KEGG enrichment analysis of 6 different genes. Figures **e** and **f** show the DO enrichment analysis of 13 genes screened



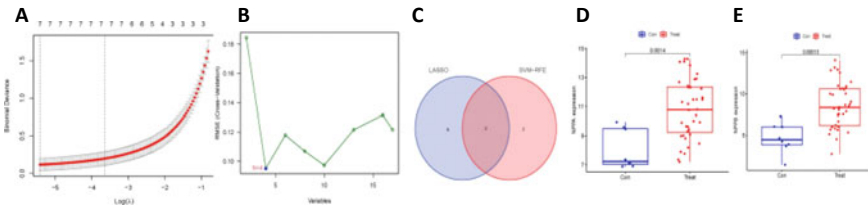
**Fig. 3** Plot of GSEA enrichment analysis of normal and DCM samples

### 3.3 Functional Clustering of DCM-GSEA Analysis

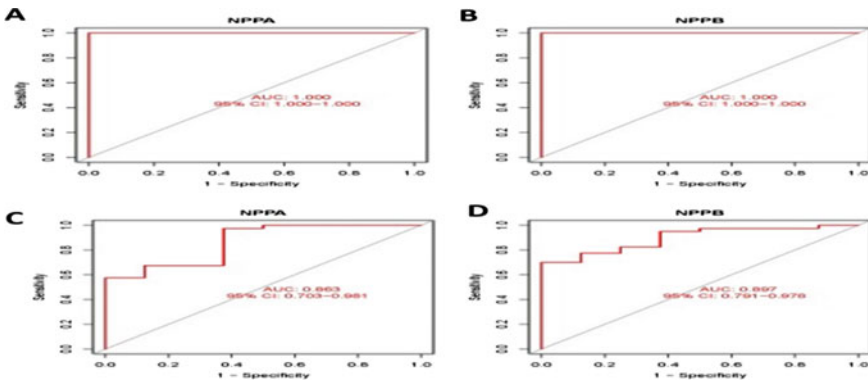
To further explore DCM-related signalling pathways, we performed GSEA enrichment analysis of GSE3585 and screened DCM-related pathways based on  $FDR < 0.05$  and adjusted  $P < 0.05$  (Fig. 3a, b). This figure shows that the following five groups of biological processes involved in immune response cell activation, gastroesophageal inflammatory arthritis response, leukocyte-mediated immunity, myeloid leukocyte activation, and mediated immunity were active in the control group; for growth factors, extracellular matrix-containing collagen, external encapsulated structures, structural components of the extracellular matrix, and structural molecular activity were mainly active in the experimental group.

### 3.4 Screening and Identification of Gene Prediction Models for Early Diagnosis

GSE3585 data were used as the training set and GSE17800 data were used as the test set. The LASSO model was constructed in the training set and the smallest value was selected for screening, which was able to obtain pivotal genes that could accurately predict early KIRC (Fig. 4a). Meanwhile, we screened 17 different genes using the SVM-RFE algorithm to obtain four hub genes (Fig. 4b). The genes obtained by these two algorithms were then intersected to obtain two key genes, NPPA and NPPB (Fig. 4c). Differential analysis crucial genes in GSE3585t revealed that the experimental group was more expressed than the normal samples (Fig. 4d, e). Subject operating characteristic (ROC) curves were constructed, and AUC values were calculated for assessing the predictive value of the model in the training and test sets. The AUCs of NPPA and NPPB in the training set were 1 (Fig. 5a, b), and the area under the curve for NPPA and NPPB in the GSE3585 group was 0.863 and 0.897 respectively, suggesting that the model has good validation performance (Fig. 5c, d).



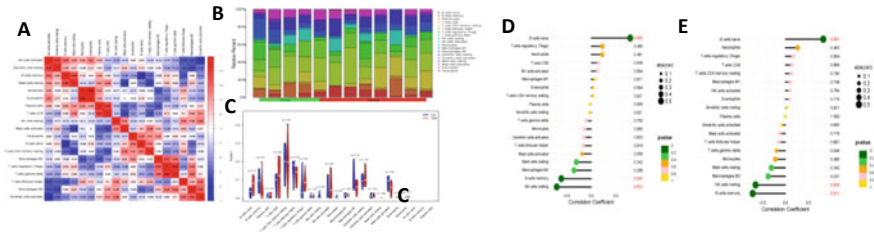
**Fig. 4** LASSO regret modelling and SVM-RFE on the training set were used to select key genes for potential DCM. **a** Lasso regression module **b** screening of key genes for potential DCM. The candidate pivotal genes of the Lasso algorithm (**c**) and 2 key gene difference analysis in normal DCM samples including NPPA (**d**), NPPB (**e**)



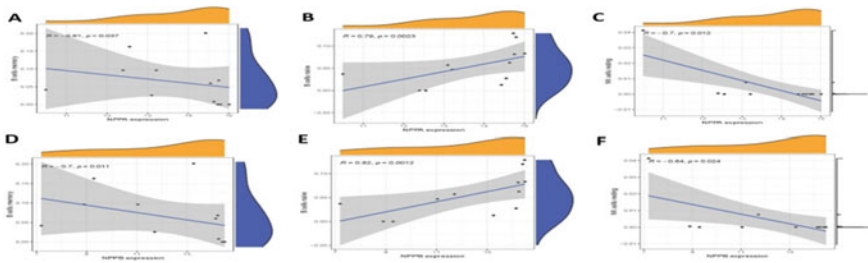
**Fig. 5** ROC curves for GSE3585 and GSE17800. NPPA (**a**), NPPB (**b**), in training set, NPPA (**c**), NPPB (**d**), in test set

### 3.5 The Immune Checkpoint Related Genes Analysis

By correlation analysis of immune cells, the results showed a positive correlation between caveolar cell quiescence and NK cell activation ( $R = 0.75$ ), mast cell resting and B cell memory ( $R = 0.69$ ), monocytes and eosinophils ( $R = 0.92$ ), T cell CD8 ( $R = -0.33$ ) and plasma cells ( $R = 0.73$ ), and T cell regulation (Tregs) and T cell  $\gamma \delta$  positive correlation ( $R = 0.84$ ) (Fig. 6a). T cell regulation (Tregs) was actively related to T  $\gamma \delta$  cells ( $R = 0.84$ ) (Fig. 6a). The results for the 12% immune cell subsets in terms of relative percentages showed that NK cells, T cells, B cells and dendritic cells accounted for the most in both the experimental and control samples (Fig. 6b). The results of the discrepancy in immune infiltration among the control and experimental groups showed that the results for each of the specimens were not significant ( $P > 0.05$ ) (Fig. 6c), which could be related to smaller sample sizes. Scatter plots were drawn by visualizing the cells with significant correlations (Fig. 7a–f), from which it can be seen that NPPA and NPPB were both negatively correlated with B cell memory ( $R = -0.61, p < 0.05$ ) ( $R = -0.7, p < 0.05$ ), NK cell resting ( $R = -0.7 p$



**Fig. 6** **a** Correlation between immune cells in GSE3585 samples. **b** Percentage of the 19 immune cell subpopulations in the two sets of samples. **c** Violin plots of the differences in immune infiltration between the two sets of samples, with normal in blue and abnormal in red. Correlation lollipop plots were drawn using key genes from the training set, including NPPA (**d**), NPPB (**e**) and immune cells



**Fig. 7** Plot of correlations between GSE3585’s essential genes and immunocells scatter plot

$< 0.05$ ) ( $R = -0.64, p < 0.05$ ) and positively correlated with B cell naïve ( $R = 0.82, p < 0.05$ ) ( $R = 0.333, p < 0.05$ ). Analysis of the linkage with key genes and vaccine cells showed that NPPA exposure was significantly linked to B-cell memory phase, NK-cell quiescence and B-cell naïve phase, all with  $P$  less than 0.05 (Fig. 6d, e).

## 4 Discussion

Dilated cardiomyopathy is a genetically predisposed myocardial disease characterised by systolic dysfunction, with or without heart failure, and its development is associated with autoimmune disorders and mutations in myosin or cytoskeletal protein genes [17]. Among these, truncated variants in the TTN gene, which encodes the giant myosin [18], as well as variants in the myosin heavy chain MYH7 and troponin TNNT2, which are involved in muscle contraction, are common genetic susceptibilities to dilated cardiomyopathy [19] and are closely related to implications for the pathogenesis of dilated cardiomyopathy. However, dilated cardiomyopathy in combination with heart failure is a clinical challenge and there are no reliable

treatment options to reverse cardiac function other than heart transplantation. There are no reliable therapeutic options to reverse cardiac function.

Based on the analysis of the GEO database, we selected two gene dataset microarrays using GPL96 and GPL570 as the study platform, with myocardial biopsies from dilated cardiomyopathy with heart failure as the experimental group and myocardial biopsies from normal heart function as the control group. Genes with up-regulated expression in the two microarrays were intersected using Venn diagrams to obtain two candidate genes. KEGG and GO analysis of the candidate genes focused on biological processes related to cell growth and conduction pathways, such as tonicity and blood clotting cascade, vascular muscle smooth muscle systole and the cGMP-PKG signalling pathway.

In this study, two key genes of NPPA and NPPB were screened, and the performance of these genes was distinct for normal and laboratory samples. DCM patients have been reported to with heart failure develop low levels of cardioselective serine proteases (corin), and serine proteases are involved in the lysis and production of active forms of natriuretic peptide precursors, so corin and natriuretic peptide family members are commonly used cardiac injury markers for clinical cardiac exhaustion, suggesting the occurrence of DCM decompensation and the severity of heart failure symptoms [19], consistent with our bioinformatic predictions. Genome-wide association studies have demonstrated that NPPA is a causative factor in blood pressure development, and that cardiac natriuretic peptide (ANP) is a vasodilating hormone encoded by NPPA that promotes salt excretion, and in humans [20], ANP levels are considered indicators of salt sensitivity [21]. Therefore, strictly limiting sodium intake and controlling hypertension may become an important way to prevent the development of DCM. At the same time, studies [22] have found that Treating DCM patients with heart failure with recombinant human brain natriuretic peptide with injection can effectively reduce the serum brain natriuretic peptide (BNP), BNP precursor (NT-proBNP), angiotensin II (AngII.) and other indexes, and improve ventricular remodeling. The crucial genes in this research are strongly linked to contributing to immunotherapy.

## 5 Conclusion

In summary, using a bioinformatics-based approach, we obtained the genes that may be related to DCM, providing meaningful research clues and directions for clinical prognosis judgment and treatment. The treatment of DCM is a personalized treatment plan, which should be further explored in clinical trials.

**Acknowledgements** This study was funded by the S&T Innovation Project for Universities in Shanxi Province (2019L0683), the Graduate Education Innovation Project in Shanxi Province (2022Y37) and the Provincial Science and Technology Grant in Shanxi Province (20210302124588).

## References

1. Merlo M, Cannatà A, Gobbo M, Stolfo D, Elliott PM, Sinagra G. Evolving concepts in dilated cardiomyopathy. *Eur J Heart Fail.* 2018;20(2):228–239. <https://doi.org/10.1002/ejhf.1103>.
2. Priori SG, Blomström-Lundqvist C, Mazzanti A, et al. 2015 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: The Task Force for the Management of Patients with Ventricular Arrhythmias and the Prevention of Sudden Cardiac Death of the European Society of Cardiology (ESC). Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC). *Eur Heart J.* 2015;36(41):2793–2867. <https://doi.org/10.1093/eurheartj/ehv316>
3. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 2013;41(D1):D991–5.
4. Wozniak MB, Le Calvez-Kelm F, Abedi-Ardekani B, et al. Integrative genome-wide gene expression profiling of clear cell renal cell carcinoma in Czech Republic and in the United States. *PLoS One.* 2013;8(3):e57886. <https://doi.org/10.1371/journal.pone.0057886>.
5. Deng YJ, Ren EH, Yuan WH, Zhang GZ, Wu ZL, Xie QQ. GRB10 and E2F3 as Diagnostic Markers of Osteoarthritis and Their Correlation with Immune Infiltration. *Diagnostics (Basel).* 2020;10(3):171. Published 2020 Mar 22. <https://doi.org/10.3390/diagnostics10030171>
6. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47. <https://doi.org/10.1093/nar/gkv007>
7. Weintraub RG, Semsarian C, Macdonald P. Dilated cardiomyopathy. *Lancet.* 2017;390(10092):400–414. [https://doi.org/10.1016/S0140-6736\(16\)31713-5](https://doi.org/10.1016/S0140-6736(16)31713-5)
8. Yu G, Wang LG, Han Y, et al. clusterProfiler: An R package for comparing biological themes among gene clusters [J]. *OMICS*, 2012, 16(5): 284–287.
9. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16(5):284–287. <https://doi.org/10.1089/omi.2011.0118>.
10. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000;25(1):25–29. <https://doi.org/10.1038/75556>.
11. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.* 2011;12:77. Published 2011 Mar 17. <https://doi.org/10.1186/1471-2105-12-77>
12. Liu, S., Xiao, Z., You, X. and Su, R., 2022. Multistrategy boosted multicolony whale virtual parallel optimization approaches. *Knowledge-Based Systems*, 242, p. 108341.
13. Huang, W., Li, Y., Zhang, K., Hou, X., Xu, J., Su, R. and Xu, H., 2021. An Efficient Multi-Scale Focusing Attention Network for Person Re-Identification. *Applied Sciences*, 11(5), p. 2010.
14. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.* 2011;12:77. Published 2011 Mar 17. <https://doi.org/10.1186/1471-2105-12-77>.
15. Yang L, Shou YH, Yang YS, Xu JH. Elucidating the immune infiltration in acne and its comparison with rosacea by integrated bioinformatics analysis. *PLoS One.* 2021;16(3):e0248650. Published 2021 Mar 24. <https://doi.org/10.1371/journal.pone.0248650>.
16. Cao Y, Tang W, Tang W. Immune cell infiltration characteristics and related core genes in lupus nephritis: results from bioinformatic analysis. *BMC Immunol.* 2019;20(1):37. Published 2019 Oct 21. <https://doi.org/10.1186/s12865-019-0316-x>
17. Walter W, Sánchez-Cabo F, Ricote M. GOplot: an R package for visually combining expression data with functional analysis. *Bioinformatics.* 2015;31(17):2912–2914. <https://doi.org/10.1093/bioinformatics/btv300>
18. Ware JS, Cook SA. Role of titin in cardiomyopathy: from DNA variants to patient stratification. *Nat Rev Cardiol.* 2018;15(4):241–252. <https://doi.org/10.1038/nrcardio.2017.190>
19. Mazzarotto F, Tayal U, Buchan RJ, et al. Reevaluating the Genetic Contribution of Monogenic Dilated Cardiomyopathy. *Circulation.* 2020;141(5):387–398. <https://doi.org/10.1161/CIRCULATIONAHA.119.037661>



20. Wood Heickman LK, DeBoer MD, Fasano A. Zonulin as a potential putative biomarker of risk for shared type 1 diabetes and celiac disease autoimmunity. *Diabetes Metab Res Rev*. 2020;36(5):e3309. <https://doi.org/10.1002/dmrr.3309>
21. Asbjornsdottir B, Snorraddottir H, Andresdottir E, et al. Zonulin-Dependent Intestinal Permeability in Children Diagnosed with Mental Disorders: A Systematic Review and Meta-Analysis. *Nutrients*. 2020;12(7):1982. Published 2020 Jul 3. <https://doi.org/10.3390/nu12071982>
22. Iwata Y, Ito S, Wakabayashi S, Kitakaze M. TRPV2 channel as a possible drug target for the treatment of heart failure. *Lab Invest*. 2020;100(2):207–217. <https://doi.org/10.1038/s41374-019-0349-z>
23. Zhu YX, Huang JQ, Ming YY, Zhuang Z, Xia H. Screening of key biomarkers of tendinopathy based on bioinformatics and machine learning algorithms. *PLoS One*. 2021;16(10):e0259475. Published 2021 Oct 29. <https://doi.org/10.1371/journal.pone.0259475>

**Others**

# Schema Based Knowledge Graph for Clinical Knowledge Representation from Structured and Un-structured Oncology Data



Farina Tariq, Saad Ahmad Khan, and Muhammad Moazam Fraz

**Abstract** Cancer is currently the second leading cause of death worldwide, killing more people every year owing to its increasing growth rate. There is a vast amount of clinical data in radiology reports and electronic health records (EHRs). Case studies are important because they offer a plethora of medical information on diseases, treatments, and other issues. However, because this information is frequently available as unstructured notes, working with it can be challenging. Additionally, the data volume is huge, the production rate is rapid, and the format is special. Thus, the conversion of health information into standards-compliant, comparable, and consistent data is essential for these scenarios. To address these challenges, we have proposed a knowledge extraction pipeline based on schema based knowledge graphs (KG), from EHRs and clinical reports. After extracting knowledge using Name Entity Recognition from radiology reports and EHRs of 33,431 cancer patients, we developed a knowledge graph in Neo4j containing 368,436 entities and 754,061 relationships of 15 different semantic categories based upon the proposed schema. The proposed method would serve as the initial step in understanding how to use KG intelligently for uniform representation of medical knowledge to analyse the course of disease after learning about it via EHRs.

**Keywords** Electronic health records · Clinical reports · Information · Extraction · Knowledge graph · Oncology data

---

F. Tariq (✉) · S. A. Khan · M. M. Fraz  
National University of Sciences and Technology (NUST), Islamabad, Pakistan  
e-mail: [ftariq.msds20seecs@seecs.edu.pk](mailto:ftariq.msds20seecs@seecs.edu.pk)

S. A. Khan  
e-mail: [skhan.msds20seecs@seecs.edu.pk](mailto:skhan.msds20seecs@seecs.edu.pk)

M. M. Fraz  
e-mail: [moazam.fraz@seecs.edu.pk](mailto:moazam.fraz@seecs.edu.pk)

# 1 Introduction

Cancer is the uncontrollable growth of body cells with one in six people losing their lives worldwide to the disease. The disease's pace of growth is predicted to climb by 70% over the next 20 years ranking it among the prevalent deadly diseases [1]. Cancer persists with no formidable solution to the ailment, resulting in un-favourable physical, emotional, and financial strains on individuals, communities, and the healthcare system. Unfortunately, many cancer types have early symptoms that show up late in the course of the disease despite biological fluids such as serum, saliva, spinal fluid, and urine being used for early detection of cancer biomarkers. Early detection of cancer may increase the likelihood of effective therapy and lower mortality rates in patients. Furthermore, they may also assist in reducing the economic burden on individuals and the health care system. Thus, the research community is actively working to achieve the aforementioned goal through means such as genomic profiling, phenotypic profiling, health disparities research and so on. The driving force behind such research is the availability of relevant data. The advent of EHRs enabled the widespread recording of digital data, both in structured and unstructured formats. Patient demographics (age, gender), height, weight, blood pressure, lab results, and medications are a few examples of structured data. Contrarily, narrative data found in EHRs such as clinical notes, surgical records, discharge summaries, radiology reports, medical photographs, and pathology reports are considered unstructured data.

The research community is interested in comprehensive cancer features which can be achieved by correlating the genomic and phenotypic data. Genomic data is found in a structured format owing to the lab panels. Phenotypic data, on the other hand, is unstructured and includes tumor morphology (such as histopathologic diagnosis), laboratory results (such as gene amplification status), particular tumor behaviors (such as metastasis), and response to treatment (such as the impact of a chemotherapeutic agent on tumor volume). This data contains vital information for analysis however the lack of structure has made it difficult to work with. Therefore, extracting phenotypic data from electronic medical records has been a top priority for many NCI-designated Cancer Centers, NCI Specialized Programs of Research Excellence, and Cancer Cooperative Groups (EMR). The extraction of cancer phenotypes is a labor-intensive, slow-moving manual process carried out by highly skilled human abstractors, making it only practical for limited datasets.

The growth of EHRs and increasing health information exchanges have resulted in the need to merge the structured and unstructured aspects of data. The integration of diverse data types across EHRs (unstructured clinical notes, time series clinical signals, static data, etc.) opens up avenues for research but is not free from its own challenges. These challenges include heterogeneous data formats (such as JavaScript object Notation (JSON), comma-separated values (CSV), and others), non-flexible storage structure (such as Relational Database Management System, or RDMS), and the lack of a big data pipeline [2]. Thus, the conversion of health information into standards-compliant, comparable, and consistent data is essential for cancer research.

The proposed framework leverages knowledge graphs to organise and manage medical knowledge that is dispersed among different EHRs components. The study has made use of the 33,431 cancer patients' clinical information that was available in the form of clinical notes and EHRs. A knowledge graph, often referred to as a semantic network, portrays a network of actual things, such as events, circumstances, or concepts, and shows how they are related to one another [3]. In addition to effectively describing and mining the relationship between medical entities and avoiding information overload, the proposed approach can describe the pertinent medical knowledge in the EHRs, clinical reports and un-structured patient notes. The proposed methodology can be used to decrease the time required for clinicians to find patient information and enhance the knowledge service capability of EHRs for improved diagnostic decisions.

## 2 Background and Related Work

There have been numerous initiatives to use healthcare data for research goals in a secondary way. In this area, numerous study efforts have been done. Richesson et al. [4] discussed the SHARPn architecture, which enables data normalisation, secure transmission, and shared phenotypic capabilities on data from many EHRs. However, they didn't employ graph-based searches as the information returned from numerous EHRs could be handled easily by the KG-based data management system. Vafajoo and his coworkers [5–7] suggested a high-performance approach that makes it easier to identify breast cancer early on using disease course analysis, before the disease progresses to the point of metastasis or the onset of tumour growth. The study has encouraged the biological, physical, and chemical discoveries for early breast cancer diagnosis.

Early studies in this area showed how to encode EHR data using an OWL model, but they did not generate RDF graphs at the patient level. The semantic web-based KG method for secondary use of cancer registry data sets has only been the subject of a small number of research. In the study proposed by Esteban-Giland [8], it suggested a semantic web-based architecture for cancer registries or the purpose of data processing and visualisation. The study's conclusions lacked clinical understanding because they were based on simulated cancer registry data rather than real data from cancer patients. By building knowledge graphs from biomedical literature, the inquiry into the risk factors for cancer and chronic illness [9] has been accomplished. The suggested process includes KG, literature-based discovery, disease-specific word embedding utilising Natural Language Processing (NLP) techniques, and Literature Based Knowledge Discovery (LBD). The KG that was developed showed that the breast cancer literature placed more stress on the clinical traits than the standard chemical prescriptions. However, this is incredibly difficult due to the vast amount and variety of biomedical data, as well as the dispersion of knowledge that is crucial for therapeutic purposes across various biomedical databases and publications. The Clinical Knowledge Graph (CKG), an open source network with more

than 16 million nodes and 220 million links that currently represents experimental data, literature, and public databases, was suggested in the study [10]. It proved that the analysis and interpretation of conventional proteomics workflows may be greatly sped up by including CKG into statistical and machine learning approaches. Several studies [11, 12] have also been carried out for relevant information retrieval from unstructured data sources like. In [13], National Cancer Institute (NCI) introduces an automated technique to extract cancer phenotype from EHRs of cancer patients in a fraction of time. The system is a double pipeline design. The first part uses ontologies for mention-extraction, and then the phenotyping summarisation pipeline uses Apache cTakes to present data in a standardised format. This system, termed DeepPhe, was compared against human expert abstracted information. The agreement between the two human experts (inter-annotator agreement) ranged from 0.46 to 1.00 (1.00 indicates perfect agreement), and system agreement with humans ranged from 0.20 to 0.96. The system, however, struggles with mentions that have multiple meanings. This could mean that the ontologies utilised were not as extensive as the authors had hoped them to be, or the manual annotation carried out had been inconsistencies with the ontology.

### 3 Design and Methodology

The goal of this work is to build disease-specific knowledge graphs in order to incorporate both structured and unstructured medical data. As shown in Fig. 1, we created the KG building framework based on EHRs by fusing the traits of tumour from EHRs. This leads to a framework consists of five steps: (1) Data description, (2) KG schema construction, (3) Information extraction, (4) Entity selection, and (5) KG construction.

#### 3.1 Data Analysis

The data which consists of radiological reports, clinical notes and EHRs of 33,431 cancer patients, which is provided by the Rawalpindi General Hospital, Pakistan.

The data is anonymized before use, and it contains EHRs with the individual clinical notes for the eight different cancer types shown in Table 1. In contrast to unstructured clinical notes, structured EHRs include patient demographic data, patients vitals, social history, diagnosis, lab results, and medication.

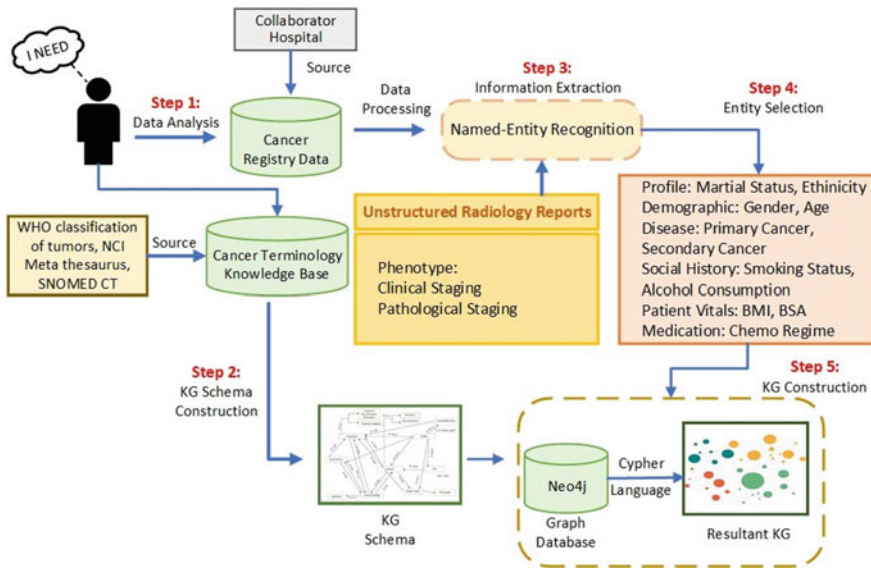


Fig. 1 Workflow of proposed methodology

Table 1 Count of patients for each cancer type

ICD-10 codes	Cancer patient
C78	2198
C79	3652
C90	1626
C77	1307
C34	4887
C18	3775
C61	5416
C50	10,570

### 3.2 KG Schema Construction

It entails examining the knowledge graph’s application demand. The knowledge graph’s goal is to provide a semantic breakdown of EHRs content at the patient level. In order to establish classes and their data qualities as well as the semantic relationships, we combined characteristics of cancer disease from many sources (SNOMED CT [14], NCI Metathesaurus [15]). Finally, the domain expert has reviewed and evaluated the schema. The relevant data is shown in Table 2.

**Table 2** Class and entities of KG schema

Class	Entities
Profile	Marital status, Ethnicity
Demographic	Age, Gender
Disease	Primary cancer, Secondary cancer
Phenotype	Clinical staging, TNM-staging, Pathological staging, Grading
Vitals	BMI, BSA
SocialHistory	Alcohol consumption, Smoking status
LabTest	Lab panel, Lab test, Lab report, Test outcome
Treatment	Chemo regime, Chemo drug, Chemo plan

### 3.3 Information Extraction

To extract relevant medical information radiology reports, the unstructured text has been subjected to a number of data preprocessing techniques: sentence splitting, co-reference resolution, abbreviation resolution and sentence simplification. The National Cancer Institute Thesaurus (NCIT) [15] ontology has been augmented with the help of domain experts from National Institute of Blood Diseases, Karachi, Pakistan to incorporate the phenotypes for both primary and secondary cancers. This ontology has been used as the knowledge base to annotate the clean notes on Inception tool. 400 notes have been manually annotated and validated by domain experts. Once the annotation has been carried out, spaCy Named-entity recognition (NER) model has been trained on the corpus for automated extraction of data in structured format from the EHRs. NER is a sub-task of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organisations, locations, expressions of times, quantities, monetary values, percentages, etc. Hand-crafted grammar-based systems typically obtain better precision, but at the cost of lower recall and months of work by experienced computational linguists.

### 3.4 Entity Selection

In medical records, the same entity may be referred to by different terms. In order to convert the original term into a standard one and construct additional entities by inheriting terms from the standard one, entity selection is therefore necessary. Our KG defines widely used sorts of entities listed by Table 2, and each of them is discussed below.

- **Profile:** The patient's ethnicity and marital status.



- **Demographic Info:** Age subdivided into three groups: less than 40 years, between 40 and 60 years, and more than 60 years where as two values of gender: male and female.
- **Disease:** International Classification Codes of Diseases (ICD-10).
- **Phenotype:** Physical or observable characteristics associated with cancer.
- **Social History:** Two factors; alcohol usage and smoking status.
- **Patient Vitals:** Body Mass Index (BMI) and Body Surface Area (BSA).
- **Lab Test:** A laboratory examination often includes numerous test items.
- **Medication:** The chemotherapy regimen with followed cycle.

### 3.5 *KG Construction*

In this research, the Neo4j graph database [16] was employed to generate the KG. Neo4j's speed is unaffected by data capacity and features a reasonably straightforward Cypher syntax. The column data types are inherited from the raw data, and we have utilised CSV files as the table name. Database column names are generated based on the column headers contained in the CSV extract. The null values in the CSV file are all set to the database null value and are displayed in a variety of ways (for example, NA, NR, and space). We have used the Cypher language to write the script for creating entity nodes from a database, as well as the relationship between these entities via linking keys.

## 4 Results and Discussion

### 4.1 *Name Entity Recognition*

The OpenNLP multi-token sequence classifier used for NER has shown an accuracy of 75.4% for the extraction of document level phenotypes. The precision and recall of the model are 0.5 and 0.4 respectively. The model has been given a note which is not annotated, and asked to identify the phenotypes without informing the model which type of cancer it is. One such example is shown in Fig. 2.

### 4.2 *EHRs Visualization by Semantic Retrieval Approach*

From a “text-centered” retrieval model to a “things-centered” retrieval approach, the knowledge graph executes the change. A multi-relational KG containing a total of 368,436 entities and 28 categories, as well as 754,061 quadruplets and 15 types, was

1	Arkansas Cancer Institute Omar T.
2	Atiq, M.D.
3	FACP 7200 South Hazel, Pine Bluff, Arkansas 71603 Phone: 870-535-2800 Fax: 870-535-2801 TestLastName TestFirstName 03/14/1968 female AtiqOmar 06/18/2019 11:30AM REQUESTING PHYSICIAN: Michelle Eckert, M.D.
4	Diagnoses: Malignant neoplasm of upper-outer quadrant of left female breast - C50.412 Essential hypertension - I10 REASON FOR CONSULTATION: Breast cancer.
5	HISTORY OF PRESENT ILLNESS: This 51-year-old female noted a mass in her left breast after a fall in January of 2019.
6	She went to see Dr.
7	Coleman, her primary care physician who confirmed physical finding and ordered a mammogram that showed the suspicious mass in the upper outer quadrant of the left breast measuring 4.3 cm in size.
8	An ultrasound confirmed it.
9	The patient then saw Dr.
10	Eckert, who did an excisional biopsy showing a 6.5 cm grade-III invasive ductal carcinoma of the breast.

Fig. 2 NER result

Table 3 Number of entities for constructed KG

Entity name	Number of entities
Profile	6588
Disease	3729
Phenotype	1168
Vitals	5687
SocialHistory	5135
LabTest	278,120
Treatment	63,180

produced via the KG synthesis framework. Each node has a special label called a property that identified it. The link between these things is described by the quadruplets. The statistics for nodes are displayed in Table 3.

In this study, we show examples of constructed KG visualisation at both the high-level and low-level. Our high-level visualisation, which is a partial visualisation of the Cancer registry data in our KG, is shown in Fig. 3. The number of edges that are connected to the node, or node degree, is the basis for the visualisation. The basic data (profile, age groups, disease, primary cancer, chemotherapy regimen, and therapeutic plan) from the EHRs are shown alongside results that have been registered. For visual considerations, only the top 300 nodes are shown. 40 and older is the key age range for enrollments. The bulk of the groups investigated are both single and married, with targeting solely non-Hispanic populations. Out of all the trials, five have combined the chemotherapy regimens of docetaxel and cisplatin (DC), cisplatin-5FU, ACDOC, and carboplatin-VP16, while the remaining studies use only chemotherapeutic agents such as Bevacizumab, Carboplatin and Casodex.

The low level visualisation of the case query for outcome data is displayed in Fig. 4. Using the phrase “Chemo regime Failure” as an example, we query the triplets (subjected-relation-object) based upon the relationship the query ‘PLAN IS DISCONTINUED’, obtaining details about the continuation of the chemo plan. 20 chemo plan nodes are collected by the results. Nine clinical chemo regimes have been

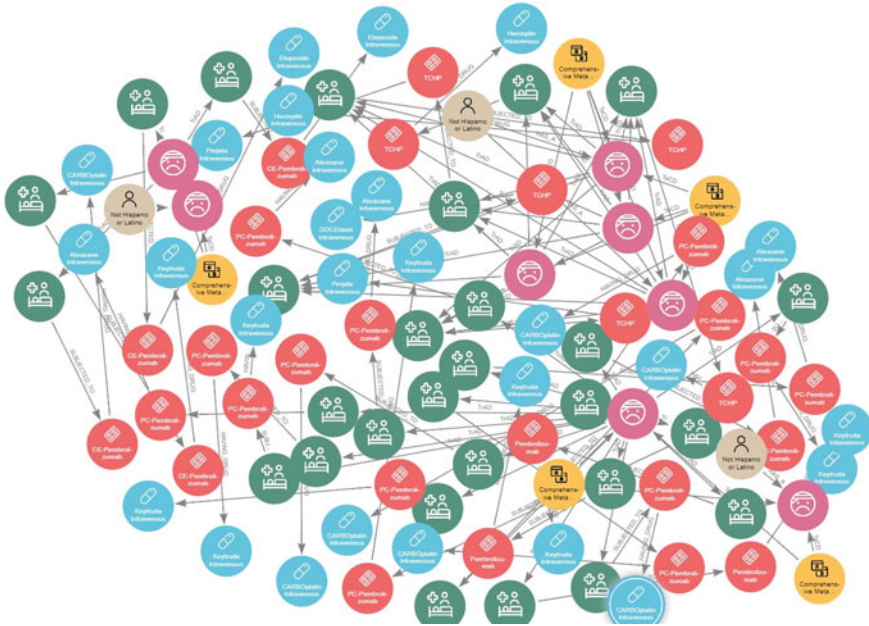


Fig. 3 High level context knowledge graph

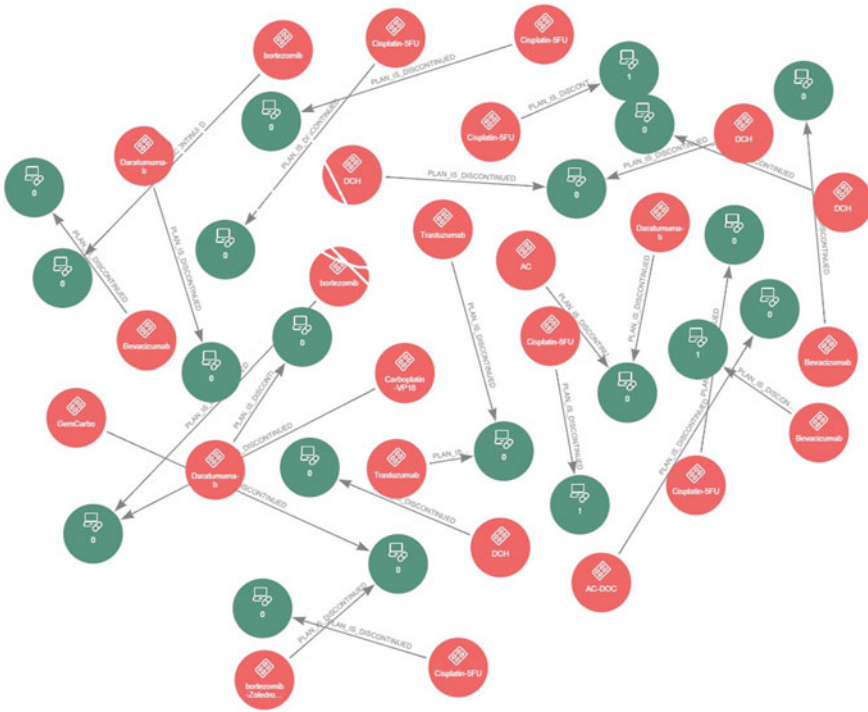
identified by the examined query, according to the extension of the node-related association. It implies that a single chemotherapy regimen might be used across several procedures.

Cypher query command:

```
MATCH p = ()-[r:PLAN IS DISCONTINUED]->() RETURN p LIMIT 25
```

## 5 Conclusion and Future Work

EHRs have a multitude of clinical knowledge, but their usage is incredibly low. The name “knowledge graph” was coined since this material is typically kept in graph databases and shown as a graph structure. A brand-new category of knowledge representation technology called KG provides an innovative method for using EHRs and deep mining. Through the use of knowledge graph approaches, we have developed a pipeline to directly combine structured data and unstructured language for application in clinical decision support systems. The learnt patient representation comprises of medication, diagnosis information, lab test, vital signs from EHRs and cancer phenotype from unstructured radiology records. A patient’s health status can be more accurately represented by this concatenation of latent representations of structured and unstructured data. Thus, it not only facilitates the organisation of



**Fig. 4** KG result of query 25 chemo-regimes with the relationship providing the information regarding chemo plan continuation

knowledge in the field of cancer research but also lays the way for knowledge extraction from unstructured radiology reports based on features. More importantly, this study promotes the development of semantics-focused ontology research. In future, by adding more instances of each phenotypes we will also enhance the performance of the NER model. The absence of a thorough assessment of the built KG is another drawback of this study. So that the KG can be utilised and assessed more extensively, we plan to develop an interactive knowledge analysis based on the created KG in the future. To achieve this purpose, We intend to use graph pattern mining methods in subsequent studies to produce clinical hypotheses drawn from this knowledge graph that improve comprehension of the patient care trajectory.

### References

1. Sung, Hyuna, et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." *CA: a cancer journal for clinicians* 71.3 (2021): 209–249.
2. Evans, R. Scott. "Electronic health records: then, now, and in the future." *Yearbook of medical*

- informatics 25.S 01 (2016): S48-S61.
3. Fensel, Dieter, et al. "Introduction: what is a knowledge graph?" *Knowledge Graphs*. Springer, Cham, 2020. 1–10.
  4. Rea, Susan, et al. "Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPN project." *Journal of biomedical informatics* 45.4 (2012): 763–771.
  5. Vafajoo, Atieh, Reza Salarian, and Navid Rabiee. "Biofunctionalized microbead arrays for early diagnosis of breast cancer." *Biomedical Physics and Engineering Express* 4.6 (2018): 065028.
  6. Ali Zaidi, Syed S., et al. "A multiapproach generalized framework for automated solution suggestion of support tickets." *International Journal of Intelligent Systems* 37.6 (2022): 3654–3681.
  7. Khurram, Imran, et al. "Dense-captionnet: a sentence generation architecture for fine-grained description of image semantics." *Cognitive Computation* 13.3 (2021): 595–611.
  8. Esteban-Gil, Angel, Jesualdo Toma's Ferna'ndez-Breis, and Martin Boeker. "Analysis and visualization of disease courses in a semantically-enabled cancer registry." *Journal of biomedical semantics* 8.1 (2017): 1–16.
  9. Daowd, Ali, et al. "A Framework To Build A Causal Knowledge Graph for Chronic Diseases and Cancers By Discovering Semantic Associations from Biomedical Literature." 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI). IEEE, 2021.
  10. Santos, Alberto, et al. "Clinical knowledge graph integrates proteomics data into clinical decision-making." *bioRxiv* (2020).
  11. Savova, Guergana K., et al. "DeepPhe: a natural language processing system for extracting cancer phenotypes from clinical records." *Cancer research* 77.21 (2017): e115–e118.
  12. Khurram, Imran, Muhammad Moazam Fraz, and Muhammad Shahzad. "Detailed sentence generation architecture for image semantics description." *International Symposium on Visual Computing*. Springer, Cham, 2018.
  13. Zhou, Sicheng. "Extracting phenotypes of cancer patients from Electronic Health Records." 2019 IEEE international conference on healthcare informatics (ICHI). IEEE, 2019.
  14. SNOMED CT. <https://www.nlm.nih.gov/healthit/snomedct/index.html>
  15. NCI Thesaurus <https://www.nlm.nih.gov/healthit/snomedct/index.html>
  16. Miller, Justin J. "Graph database applications and concepts with Neo4j." *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*. Vol. 2324. No. 36. 2013.

# Intelligent Fuzzy Clinical Decision Support System to Classify Breast Cancer—Case Study: The Wisconsin Dataset



Y. F. Hernández-Julio , L. A. Díaz-Pertuz , M. Prieto-Guevara ,  
M. Avilés-Román , B. Castillo-Osorio , M. Barrios-Barrios ,  
and W. Nieto-Bernal 

**Abstract** The aim of this research work was to design, implement, and validate several fuzzy decision support systems using clusters and dynamic tables. The outcomes were evaluated with related works obtained from the literature for validating the proposed fuzzy inference systems—FISs to classify the Wisconsin breast cancer dataset. Two validation approaches were used in this work. The FISs were trained with distinct input features, according to the studies obtained from the literature. The uniqueness of this study remains in the manner of generating the membership functions and the rule base for the intelligent fuzzy clinical decision support systems. The outcomes showed that the obtained performance metrics in several issues were higher than the achieved outcomes from the literature, demonstrating better precision.

**Keywords** Fuzzy system · Breast cancer · Clusters · Pivot tables

---

Y. F. Hernández-Julio (✉) · L. A. Díaz-Pertuz · M. Avilés-Román · B. Castillo-Osorio  
Faculty of Economics, Administrative and Accounting Sciences, Universidad del Sinú Elías  
Bechara Zainúm, 230002 Montería, Colombia  
e-mail: [yamidhernandezj@unisinu.edu.co](mailto:yamidhernandezj@unisinu.edu.co)

M. Prieto-Guevara  
Universidad de Córdoba, 230002 Montería, Colombia

M. Barrios-Barrios  
Universidad de la Costa, 080001 Barranquilla, Colombia

W. Nieto-Bernal  
Universidad del Norte, 080001 Puerto Colombia, Colombia

## 1 Introduction

Breast cancer is the second most common cancer among women in the United States (some kinds of skin cancer are the most common [1]). According to [2], these are some symptoms of breast cancer, a lump or swell in the breast, upper chest, or armpit, among others. To classify breast cancer fuzzy logic has been used [3, 4].

For the reasons above, the aim of this study was the design, implementation, and validation of different fuzzy systems using a framework for the development of data-driven fuzzy clinical decision support systems for classification problems. To validate the proposed models, fuzzy inference systems—FIS were trained and tested for classifying the Wisconsin Breast Cancer dataset and evaluated the results with other artificial intelligence techniques models acquired from the literature.

## 2 Material and Methods

For validating the framework proposed by [5, 6], different experiments were implemented. Each of the phases that are mentioned in the framework for the development of the clinical decision support systems through clusters and dynamic tables was used.

### 2.1 *Identifying the Dataset*

The Wisconsin Breast Cancer Dataset (WBCD) [7] comprises 458 benign instances and 241 malignant instances. The properties of this dataset are explained in Ref. [8].

### 2.2 *Data Preparation (Crisp Inputs)*

In this phase, the data had to be changed because the dataset has missing values. The symbol “?” was switched by a number 0. Also, we had to change the class 2 by 1 for benign class and 4 by 2 for malignant class. In this stage, the clustering methods were applied. This step will be described in Sect. 2.7.

### ***2.3 Reviewing Existing Models***

In this stage, a search of the different related works about the topic was carried out. The indexed databases such as Scopus, Science Direct, among others were used.

### ***2.4 Evaluating the Optimal Number of Clusters***

In this stage, pivot tables were applied to the WBCD to know the number of rows for all the used variables. For the case study, the optimal number of clusters was 10.

### ***2.5 Setting a Number of Clusters (Minimum and Maximum) According to the Previous Evaluation***

The minimum values were 2, and the maximum number of clusters was ten for all input variables. To the output feature, two clusters were applied.

### ***2.6 Random Permutations***

The dataset were randomized and permuted when used the proposed algorithms.

### ***2.7 Cluster Analysis (Fuzzification Process)***

In this phase, three clusters methods were applied and analyzed using the results obtained in the previous stage. The maximum number of clusters for every variable for each subset was the values of the optimal cluster.

### ***2.8 Sampling Datasets (Cross-Validation or Random Sampling)***

For the experiments, two random sampling methods were applied. In one hand, random sampling was used, and in the other hand the cross-validation method was used. The default values for the first data partition method were training dataset: 70%, validation dataset: 30%, and number of iterations: 3000. For the cross-validation process, the k-fold method was selected. The default value of k was 10.



## **2.9 Pivot Tables**

To the experiments, the “unique” command from the Matlab software was utilized for the deployment of the subsequent sub-stages. This stage helps to establish the rules number for the development of the DDFCDSS.

### **2.9.1 Combining Different Input Variable Clusters Datasets**

This phase involves of creating arrangements among input features and the sets of output features applying dynamic tables.

### **2.9.2 Stablishing the Fuzzy Rules**

This phase is established on the earlier one. The procedures involved using the pivot tables one or several permutations can be used to make the rule bases for the FIS.

## **2.10 Elaborating the Decision Support System Based on Fuzzy Set Theory (Inference Engine)**

In this phase, the aim is to put all the elements cited above in the fuzzy inference system.

### **2.11 Evaluating the Fuzzy System Performance (Defuzzification and Crisp Outputs)**

This phase is assessing through some performance metrics (showed in Tables 2, 3, 4, 5, 6, 7, 8 and 9). Also, we performed a McNemar’s test.

## **3 Results and Discussion**

Table 1 shows the confusion matrix for the mentioned Data-Driven Fuzzy Clinical Decision Support System (DDFCDSS).

Tables 2 and 3 show the performance metrics obtained with our proposed framework. The most excellent results for a set of five features were obtained by the Ward clustering method using Random sampling validation method.

**Table 1** Confusion matrix for WCDB dataset

		Specialists	
		Benign	Malign
DDFDSS	Benign	<b>455</b>	3
	Malign	0	<b>241</b>

Bold text represents accurate forecasts  
 DDFDSS Data-driven fuzzy decision support system

**Table 2** Performance metrics achieved using the proposed framework (cross-validation method)

[5]	CV		
[2 4 5 6 8]	K-means	Ward	FCM
Num of rules or hidden neurons/technique	248	233	190
Accuracy (%)	99.3	99.4	99.1
Sensitivity	0.9857	0.9853	0.9851
Specificity	0.9969	0.998	0.9939
F-Measure	0.9899	0.9907	0.9868
Area under curve:	0.9933	0.9942	0.9903
Kappa statistics:	0.9845	0.9858	0.9798

CV cross-validation method

**Table 3** Performance metrics achieved using the proposed framework (random sampling method)

[5]	RS		
[2 4 5 6 8]	K-means*	Ward	FCM*
Num of rules or hidden neurons/technique	207	208	168
Accuracy (%)	99.0	99.57	98.43
Sensitivity	0.9916	0.9877	0.9637
Specificity	0.9892	1.0000	0.9956
F-Measure	0.9854	0.9938	0.9775
Area under curve	0.9874	0.9967	0.986
Kappa statistics	0.9778	0.9905	0.9654

RS random sampling

\*Significant difference at 95% of the Confidence Interval between them

As can be seen, the DDFCDS had a specificity value of 100%, showing an excellent performance predicting the true negatives cases of the Wisconsin Breast Cancer Dataset (WBCD). It means that all malignant cases were classified correctly. According to the confusion matrix, there are only three true positive values misclassified corresponding to a sensitivity value of 0.9877.

In the following pages, we are going to compare the results obtained from the literature with our results. The results shown in the Table 4 correspond to the same

**Table 4** Performance metrics achieved using the proposed framework compared with results obtained by [8] (cross-validation method)

Onan [8]		CV		
[1 2 4 5 6 7 8]		K-means	Ward	FCM
Num of rules or hidden neurons/technique	FRNN	312	338	260
Accuracy (%)	99.72	98.94	98.53	98.66
Sensitivity	1.0000	0.9703	0.9594	0.9694
Specificity	0.9947	1.0000	1.0000	0.9960
F-Measure	0.9970	0.9849	0.9792	0.9808
Area under curve	1.0000	0.9919	0.9888	0.9880
Kappa statistics	0.9943	0.9768	0.9679	0.9704

FRNN fuzzy rough nearest neighbor, CV cross-validation

**Table 5** Performing metrics obtained with our proposed framework compared with results obtained by [8] (random sampling method)

Onan [8]		RS		
[1 2 4 5 6 7 8]		K-means	Ward	FCM
Num of rules or hidden neurons/technique	FRNN	260	253	249
Accuracy (%)	99.72	98.66	98.00	98.28
Sensitivity	1.0000	0.9694	0.9451	0.9526
Specificity	0.9947	0.9960	1.0000	1.0000
F-Measure	0.9970	0.9808	0.9718	0.9757
Area under curve	1.0000	0.9880	0.9737	0.9869
Kappa statistics	0.9943	0.9704	0.9563	0.9624

RS random sampling, FRNN fuzzy-rough nearest neighbor

<sup>NS</sup>Not Significant difference at 95% of the confidence interval

characteristics made by the researchers using the same dataset (WBCD). We used the same data partition method, the same features.

According to the results showed in Tables 4 and 5, for the WBCD, the best performance belongs to Onan [8]. The author used a tenfold cross-validation method as data partition. As can be seen in Tables 4 and 5, the classification accuracy for his results was 99.72%, and the maximum value for classification accuracy of our results belong to the k-means tenfold cross-validation method. The obtained sensitivity value by the author was 100%; however, his specificity value was 0.9947. Our results show the contrary. Our Specificity value was 1.0, and the sensitivity value was 0.9703. The performance metric Sensitivity indicates the true positive (TP) rate, and specificity means the true negative (TN) rate [9]. According to [9], in Breast Cancer the TP represents cases that are correctly categorized in the benign tumor, and the TN represents cases that are correctly categorized in the malign tumor. This result shows that our model predicts 100% of the true negative values. In this case, we can state

that if a tumor is malignant, the fuzzy inference system is going to be classified as malignant with a 100% of accuracy.

Making the comparison between the three clustering methods results, we found that McNemar’s test indicated that all of them don’t perform significantly better than the other, indicating that all the DDFCDSS have the same classification error rates. The test results were Ward vs. k-means:  $X_1^2 = 0.0455$ ; k-means vs. FCM:  $X_1^2 = 0.0$ , and Ward vs. FCM:  $X_1^2 = 0.12903$ , respectively.

According to Tables 6 and 7, Zemouri et al. [10] proposed a Breast Cancer Computer Aid Diagnosis (BC-CAD) based on joint variable selection and a Constructive Deep Neural Network “ConstDeepNet”. The authors used fivefold cross-validation as a partition data method. The classification accuracy for the set of features mentioned in Tables 6 and 7, is 96.2%. Our results were higher than the obtained for these authors. Our classification accuracy using Cross-validation data partition method with  $k = 5$  was 98.37%. For the comparison among the three clustering approaches, the McNemar’s test outcomes are the following: K-means vs. Ward:  $X_1^2 = 0.3636$ ; k-means vs. FCM:  $X_1^2 = 1.8947$ , and Ward vs. FCM:  $X_1^2 = 0.5625$ , indicating no significant differences between them. For the case of the second set of features used by the authors (Tables 8 and 9), the obtained classification accuracy by the constructive deep neural network was 96.6%. Our results for the same set of features were higher than the obtained by the authors. Regarding the McNemar’s test results for the three clustering methods, they indicate that there is no significant difference among them. The test values are k-means vs. Ward:  $X_1^2 = 0.3636$ ; k-means vs. FCM:  $X_1^2 = 1.8947$ ; Ward vs. FCM:  $X_1^2 = 0.5625$ .

**Table 6** Performance metrics achieved using the proposed framework compared with results obtained by [10] (cross-validation method)

[10]		CV		
[1 4 5 6 8 9]		K-means	Ward	FCM
Num of rules or hidden neurons/technique	DNN	198	214	199
Accuracy (%)	96.2	98.37	98.26	98.31
Sensitivity	–	0.9713	0.9736	0.9621
Specificity	–	0.9903	0.9873	0.9947
F-Measure	–	0.9765	0.9747	0.9759
Area under curve	–	0.9832	0.981	0.9848
Kappa statistics	–	96.40%	96.14%	96.29%

DNN deep neural network, CV cross-validation

**Table 7** Performance metrics achieved using the proposed framework compared with results obtained by [10] (random sampling method)

[10]		RS		
[1 4 5 6 8 9]		K-means <sup>NS</sup>	Ward <sup>NS</sup>	FCM <sup>NS</sup>
Num of rules or hidden neurons/technique	DNN	192	193	180
Accuracy (%)	96.2	99.00	98.86	98.86
Sensitivity	–	0.9875	0.9915	0.9794
Specificity	–	0.9913	0.9870	0.9934
F-Measure	–	0.9854	0.9832	0.9835
Area under curve	–	0.9884	0.9844	0.9883
Kappa statistics	–	97.8%	97.45%	97.47%

RS random sampling. – not mentioned in the literature, DNN deep neural network

<sup>NS</sup>Not significant difference at 95% of the confidence interval

**Table 8** Performance metrics achieved using the proposed framework compared with results obtained by [10] (cross-validation method)

[10]		CV		
[1 2 5 6 7 8]		K-means	Ward	FCM
Num of rules or hidden neurons/technique	DNN	212	221	147
Accuracy (%)	96.6	98.63	98.51	96.25
Sensitivity	–	0.9684	0.9684	0.959
Specificity	–	0.996	0.9943	0.9646
F-Measure	–	0.9803	0.9787	0.9448
Area under curve	–	0.9878	0.9861	0.9552
Kappa statistics	–	96.98%	96.73%	91.65%

DNN deep neural network. CV cross-validation

**Table 9** Performance metrics achieved using the proposed framework compared with results obtained by [10] (random sampling method)

[10]		RS		
[1 2 5 6 7 8]		K-means <sup>NS</sup>	Ward <sup>NS</sup>	FCM <sup>NS</sup>
Num of rules or hidden neurons/technique	DNN	183	198	178
Accuracy (%)	96.6	99.00	98.86	98.00
Sensitivity	–	0.9875	0.9794	0.9595
Specificity	–	0.9913	0.9934	0.9912
F-Measure	–	0.9854	0.9835	0.9713
Area under curve	–	0.9884	0.9883	0.9808
Kappa statistics	–	97.8%	97.5%	95.6%

RS random sampling, – not mentioned in the literature, DNN deep neural network

## 4 Conclusions

The objective of this research work was the design, implementation and validation of the diverse decision support systems based on a fuzzy set theory applying clustering methods and pivot tables. As could be demonstrated, in some cases, the proposed fuzzy models showed the best performing indices related to this dataset, surpassing the outcomes obtained from advanced artificial intelligence techniques (deep learning). The achieved outcomes for the used metrics were closer to 100%, indicating a strong fit between the forecast results and the studied data. The obtained performance metrics for this dataset were between 0.90 and 1.0, representing an excellent classification task [11]. The chosen features shown in Tables 2 and 3 using both data partition methods were: Uniformity of Cell Size, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Normal Nucleoli, In this case, it is not essential to perform the process of mitosis, reducing waiting times for the decision-making, accelerating a probable treatment [5, 12]. According to the McNemar's test results for the three clustering methods, the k-means have significant difference at 95% of the confidence interval with the FCM clusters method ( $X_1^2 = 5.7857$ ), indicating that these two clusters' methods have different error rate. For the other two clusters methods the test evidenced that the clustering methods did not perform significantly different.

## References

1. Reis, H.C.; Turk, V.; Khoshelham, K.; Kaya, S. InSiNet: a deep convolutional approach to skin cancer detection and segmentation. *Medical & Biological Engineering & Computing* **2022**, *60*, 643–662, doi:<https://doi.org/10.1007/s11517-021-02473-0>.
2. Breast Cancer Now. What are the signs and symptoms of breast cancer? Available online: <https://breastcancernow.org/about-us/media/facts-statistics#signs-and-symptoms> (accessed on February).
3. Nilashi, M.; Ibrahim, O.; Ahmadi, H.; Shahmoradi, L. A knowledge-based system for breast cancer classification using fuzzy logic method. *Telematics and Informatics* **2017**, *34*, 133–144, doi: <https://doi.org/10.1016/j.tele.2017.01.007>.
4. Gayathri, B.M.; Sumathi, C.P. Mamdani fuzzy inference system for breast cancer risk detection. In Proceedings of the 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICIC), 10–12 Dec. 2015, 2015; pp. 1–6.
5. Hernández-Julio, Y.F.; Prieto-Guevara, M.J.; Nieto-Bernal, W.; Meriño-Fuentes, I.; Guerrero-Avenida, A. Framework for the development of data-driven Mamdani-type fuzzy clinical decision support systems. *Diagnostics* **2019**, *9*, 1–33.
6. Hernández Julio, Y.F.; Nieto Bernal, W.; Muñoz Hernández, H. *Framework for the development of data-driven mamdani-type fuzzy decision support systems based on fuzzy set theory using clusters and pivot tables*; Ed. Universidad del Sinú Elías Bechara Zainúm: Montería, Córdoba, Colombia, 2021; Volume 1, p. 108.
7. Bache, K.; Lichman, M. UCI machine learning repository. **2013**, [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(original\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)).
8. Onan, A. A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. *Expert Systems with Applications* **2015**, *42*, 6844–6852, doi: <https://doi.org/10.1016/j.eswa.2015.05.006>.

9. Liu, K.; Kang, G.; Zhang, N.; Hou, B. Breast Cancer Classification Based on Fully-Connected Layer First Convolutional Neural Networks. *IEEE Access* **2018**, *6*, 23722–23732.
10. Zemouri, R.; Omri, N.; Devalland, C.; Arnould, L.; Morello, B.; Zerhouni, N.; Fnaiech, F. Breast cancer diagnosis based on joint variable selection and Constructive Deep Neural Network. In Proceedings of the 2018 IEEE 4th Middle East Conference on Biomedical Engineering (MECBME), 28–30 March 2018, 2018; pp. 159–164.
11. Gorunescu, F. *Data Mining: Concepts, Models and Techniques*; Springer-Verlag Berlin Heidelberg: Berlin, 2011; Volume 12, p. 372.
12. Hernández-Julio, Y.F.; Hernández, H.M.; Guzmán, J.D.C.; Nieto-Bernal, W.; Díaz, R.R.G.; Ferraz, P.P. Fuzzy Knowledge Discovery and Decision-Making Through Clustering and Dynamic Tables: Application in Medicine. In *Information Technology and Systems. ICITS 2019. Advances in Intelligent Systems and Computing*, Rocha, Á., Ferrás, C., Paredes, M., Eds.; Springer, Cham: Quito, Ecuador, 2019; Volume 918, pp. 122–130.

# Research on the Design and Production of VR Rehabilitation Game for Parkinson's Disease Patients Based on Real-Time Action Acquisition



Ying Zhang, Xin Su, and Xibin Xu

**Abstract** Aiming at the symptoms such as motor retardation of Parkinson's patients, this paper studies real-time motion acquisition based on Kinect equipment, and designs remote virtual reality (VR) rehabilitation games using Unity as well as visual studio programming tools to help patients perform interesting rehabilitation movement exercises. It also stores the game score data obtained through the relevant database platform, forming the patient rehabilitation training results. Finally, this result will be used as the basis for the evaluation of VR rehabilitation games and serves as a long-term rehabilitation auxiliary treatment mechanism for Parkinson's patients.

**Keywords** Virtual reality · Parkinson's · Rehabilitation games

## 1 Introduction

### 1.1 Rehabilitation Treatment of Parkinson's Disease

Parkinson's disease [1] is a disease based on neuropathy, and the onset age of most people is about 60 years old. Parkinson's patients have the symptoms of dyskinesia, which is due to the loss of dopaminergic neurons in the substantia neuropathology of the patient's midbrain. It will lead to the reduction of dopamine production in

---

This work was supported by DATANG MOBILE COMMUNICATIONS EQUIPMENT CO., LTD (20202000659), and by the National Key R&D Program of China under Grant (2020YFB1806702).

---

Y. Zhang

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China

X. Su (✉) · X. Xu (✉)

Department of Electronic Engineering, Tsinghua University, Beijing, China  
e-mail: [suxin@tsinghua.edu.cn](mailto:suxin@tsinghua.edu.cn)

X. Xu

e-mail: [xuxb@tsinghua.edu.cn](mailto:xuxb@tsinghua.edu.cn)



the midbrain, resulting in increased muscle tone, tremor, slow movement, and dyskinesia. China has the largest number of Parkinson's patients in the world. According to the prediction of experts from the World Health Organization, about 500,000 people in China will suffer from this disease by 2023. Although there are about 100,000 new patients every year, the medical treatment rate in China is extremely low, accounting for only 40% of them. Only when patients with Parkinson's disease carry out appropriate activities can their dyskinesia symptoms be effectively alleviated. They need to choose appropriate sports according to their own conditions or doctor intervention, such as table tennis and swimming for patients who are not too old. Older patients can do yoga and catch butterflies slowly. Only when patients choose appropriate rehabilitation exercise, can it be carried out correctly and normatively, and combined with doctors' suggestions, they can get twice the result with half the effort. At present, Parkinson's disease is not completely curable. Once the elder suffer from this disease, they generally need to take active rehabilitation exercises. In view of the traditional "one-to-one" rehabilitation mechanism for Parkinson's disease patients, rehabilitation practitioners will not be able to meet the doubled number of elder patients, which will lead to the shortage of health care workers for rehabilitation training, and greatly reduce the rehabilitation efficiency. At the same time, elder patients not only have time and space limitations that will bring inconvenience to the tracking and feedback of medical staff and the follow-up of their conditions, but also from the perspective of patients, the long-term quantitative training content is boring, which is easy to put pressure on the spirit and psychology of elder patients and bring economic stress. Therefore, it is particularly important for the elder patients with Parkinson's disease to achieve the rehabilitation effect through virtual reality game training.

## ***1.2 Rehabilitation Therapy with Kinect***

In recent years, Kinect, a depth camera, has been widely used in the medical field [2–5]. Many scientific research institutions at home and abroad are using it to carry out relevant research, develop sports games with Kinect, and help patients with movement disorders to carry out rehabilitation treatment. The Royal Berkshire Hospital in the UK [6] used the game developed by Kinect to help stroke patients regain their athletic ability. Southampton University [3] also developed a technology based on Kinect to help stroke patients recover. The Hong Kong Polytechnic University [7] developed kinelabs somatosensory games to train the sports ability of stroke patients and help them recover. A company in Hangzhou has developed an "intelligent motion analysis and training system" [8] that converts the control of human motion into the supervision of limb motion. It has been released and put into clinical practice in Zhejiang Provincial People's Hospital and other hospitals. Fahmy [9] and others have successfully developed a shoulder rehabilitation system in their research, which improves the rehabilitation speed and reduces the possibility of secondary injury by setting corresponding parameters in combination with various situations.

### ***1.3 Effects of VR Rehabilitation Games on Parkinson's Patients***

VR technology acts on users through perception. With the help of necessary equipment, the user [10] maps himself into the virtual scene, communicates with the objects in the scene, and makes them experience their surroundings, thus making the feeling of human-computer interaction more real. Today, VR technology has been applied to all aspects of rehabilitation treatment, similar to the rehabilitation of motor disorders in upper limb imbalance, cognitive rehabilitation in attention focus, and emotional rehabilitation in depression, all of which have achieved good rehabilitation effects [11]. Among them, rehabilitation training of motor dysfunction is one of the important contents of rehabilitation medicine [12, 13]. Based on the control experimental research, it has been proved that by selecting appropriate games as references and integrating safety and intensity according to the characteristics of medical rehabilitation assistance needs, gait balance and cognitive functions of Parkinson's patients can be improved more than traditional gait training.

### ***1.4 VR Rehabilitation Games can be Played Remotely with the Help of the Network***

With the rapid development of 5G technology, some problems such as slow loading caused by data transmission quality in the past have been fully delayed. Zhejiang Mobile and Wenzhou central blood station have made concerted efforts to build an "information highway" that applies 5G to free blood donation for smart blood stations. By providing 5G+ VR panoramic experience in blood donation rooms and other areas, when blood donors wear VR glasses that experience every brilliant moment from an all-round perspective, they reduce their tension when donating blood. The ultra bandwidth transmission of 5G not only effectively solves the problem of strong dizziness caused by the use of helmets, and makes them more relaxed in the process of blood donation, but also narrows the psychological distance between doctors and blood donors. In addition, Hangzhou people's Hospital also released the VR psychological rehabilitation diagnosis and treatment system based on the clinical project of 5G technology. With the arrival of 5G era, telemedicine came into being, and the distance between patients and hospitals has been greatly reduced. In addition to remote surgery and video consultation becoming research hotspots in the medical field, rehabilitation medicine will also become an indispensable part of the follow-up development. 5G technology undoubtedly provides timely assistance for the development of VR. In terms of data transmission, while 4G offered the maximum data rate of 150 Mbps per user under ideal conditions, 5G promises 10Gbps under the same conditions. In terms of delay, compared with 100 ms for 3G and 30 ms for 4G, 5G promises 1 ms. In connection density, 4G could connect 100 thousand devices per square kilometer, while 5G promises to connect 1 million per square

kilometer. In a word, the above advantages help the panoramic virtual reality scene in the telemedicine game that need higher resolution and transcoding rate.

## 2 Virtual Reality Rehabilitation Game Design

### 2.1 System Architecture of Butterfly Catching Game

The overall structure of the game is shown (see Fig. 1). The basic idea of the rehabilitation game is that the patient stands on the grass, and there are many butterflies around him at random. After catching butterflies with his fingers for 5 s, the butterflies disappear, and the number of butterflies he catches is calculated. The game logic mainly controls the number of butterflies, simulates the flying of butterflies, and detects the realization of catching butterflies.

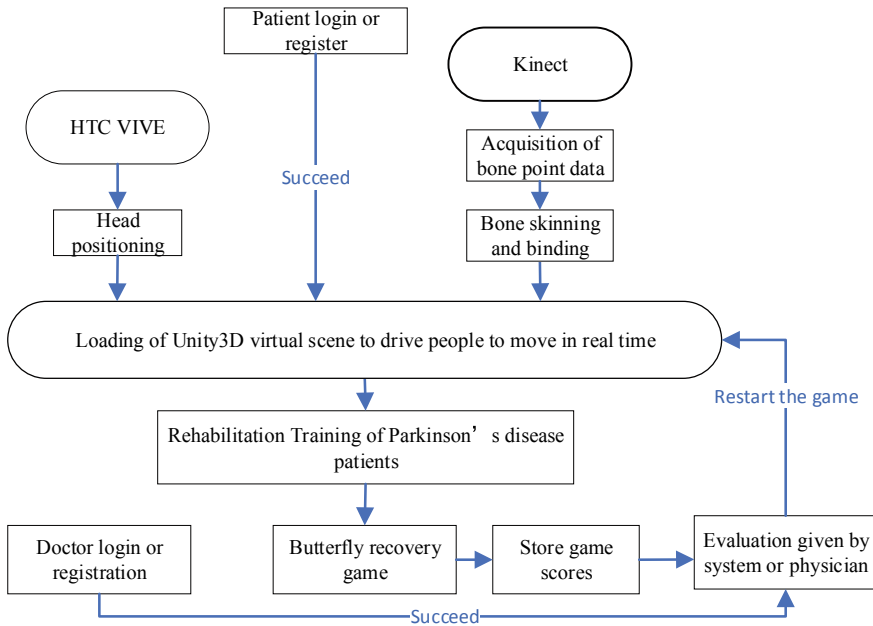


Fig. 1 System architecture of butterfly catching game

## 2.2 Control the Number of Butterflies

Use the for loop to traverse the butterflies in the scene. When the number of butterflies set is less than the existing butterflies, the cloning function will be automatically triggered. First, the butterfly is made into a preform, and then the preform is cloned using the “instantate” function.

## 2.3 Function Design for Simulating Butterfly Flying

In order to achieve the goal of simulating butterfly flying, customize butterfly flying and rest parameters. Judge whether the butterfly model is empty. If not, get the flying speed of the butterfly component and turn off the rest parameters.

## 2.4 Realization of Detecting and Catching Butterflies

The realization of catching butterflies is based on collision detection. Any collision detection has two carriers, one is to initiate the collision and the other is to accept the collision. The collision effect is like that the protagonist is a rigid body and the hand is a collision body. The protagonist will disappear after touching the hand. First, add rigid bodies to the fingers of the butterfly preform and manikin. Then we add a collision body to the butterfly, and a script to the character to monitor how long the character has been touched by which butterfly.

## 2.5 Design of Database

The design of the rehabilitation game table is shown in Table 1. This task is to create a simple login and registration interface. The system can determine whether the user already exists when clicking the login button, and if so, log in directly and load the user’s data. If the user does not exist, it will prompt login failure. When the user clicks

**Table 1** Butterfly catching system database table

Field	Field type	Remarks
User-id	Int(8)	Primary key
Username	Varchar(32)	Foreign key
Password	Varchar(32)	Foreign key
Gender	Char(4)	Foreign key
Score	Int(8)	Foreign key

the register button, it will judge whether the user already exists. If the user already exists, the system will prompt “registration failed, the user name already exists”. If the user does not exist, it will prompt “registration succeeded”. After the user logs in successfully, the game scene is loaded. In the game scene, it can realize that every time you catch a butterfly, the butterfly will disappear and the score will be increased by one.

### 3 Rehabilitation Game Making Based on Virtual Reality

#### 3.1 Unity3D Components

Each complete Unity3D game is composed of several game scenes, game objects and game components. The game scene is composed of several game objects, and the game objects are composed of several components. All game objects can be created on the hierarchy panel, and all game resources can be managed in the project view. Components are also called scripts. There are some parameters on each component that can be selected or designed. For example, the transform component is available to all game objects and cannot be deleted. It is used to record the coordinates, rotation and zooming information of game objects in the 3D world.

#### 3.2 Scene Production of Rehabilitation Games

The login scenario of the game is shown (see Fig. 2). After the patient logs in, it will automatically shift to the main scene of the game, as shown (see Fig. 3). When

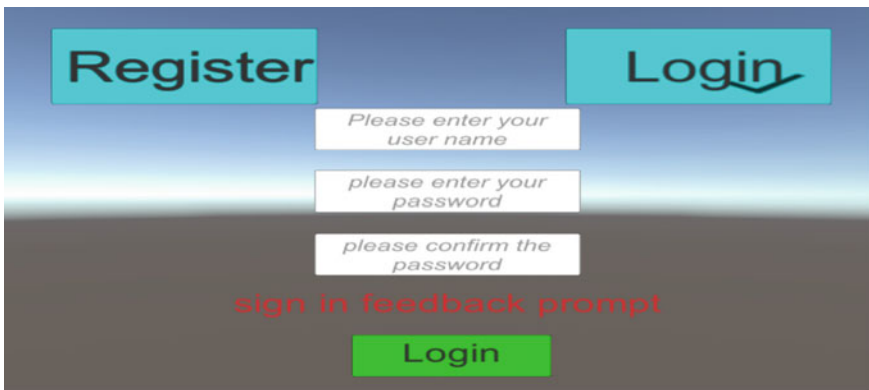
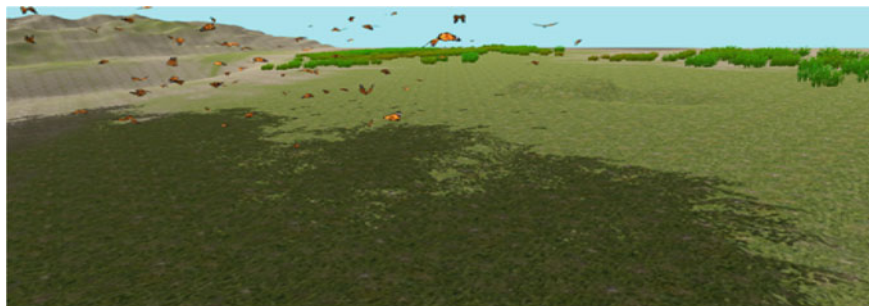


Fig. 2 Login page of the game



**Fig. 3** Scene of the game of catching butterflies

patients do rehabilitation training in an open room for a while, they will inevitably get bored. Therefore, starting from the patient's experience and use effect, this system focuses on improving the delicacy of face-to-face pictures and interactions of patients during use, such as beautiful scene building and model rendering, and is committed to improving the patient's experience during long-term rehabilitation. In the rehabilitation game, this paper uses the terrain production of Unity3D to simulate the wild butterfly flying environment and feel the beauty of nature. The inconsistent direction of butterflies increases the spontaneity of patients' movement and allows them to enjoy a comfortable time.

### ***3.3 Scene Production of Rehabilitation Games***

The patient reaches the virtual environment by using HTC VIVE, a device that can read the user's position coordinates in space. Then kinect collects depth information 7 through infrared information, recognizes other people's body parts according to machine learning algorithms, and thus obtains 25 bone point coordinates. Kinect is used to capture human actions in real time, and the captured bone information is transmitted to Unity in real time. Unity assigns the obtained data processing to the character model in the demo, so as to realize real-time synchronization of human actions to virtual characters. The specific character animation model is shown (see Fig. 4). The connection process between Kinect and Unity is as follows: First, install Kinect SDK on your computer. Then, in the Unity3d project, import the Carnegie Mellon plug-in. Create an empty object in the scene and hang the Kinect manager script on it. Finally, drag the "avatarcontroller" script onto the character model. In order for the virtual character model to keep up with the rhythm of the actual character, you need to drag and drop the key skeleton points on the model that control the movement of the model to the appropriate variables in the control script, that is, bind the skeleton points recognized by Kinect to the skeleton points in the model. Part of the skeleton points of the model are shown (see Fig. 5). After binding



Fig. 4 Character animation model

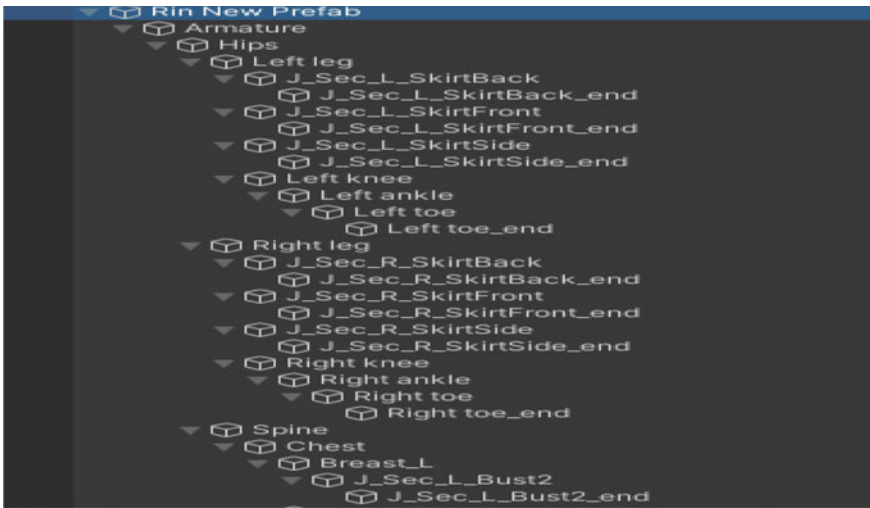


Fig. 5 Some bone points in the character animation model

the skeleton points, you can use Kinect to control the movement of the characters in Unity.

## 4 Rehabilitation Game Making Based on Virtual Reality

In this paper, the design and production of VR rehabilitation games for Parkinson's disease patients are described, and in combination with the rapid development of 5G wireless communication, it is proposed that Parkinson's disease patients can

be rehabilitated through remote VR rehabilitation games to achieve the purpose of reducing movement disorders. At the same time, the design and implementation of the main functions of the system, as well as the relevant manufacturing technologies are briefly described. In order to achieve a better immersion effect, this paper analyzes and compares the current game development engines such as Unity3d game engine, and completes the design and production of butterfly catching VR rehabilitation game based on Kinect. The running effect shows that the VR rehabilitation game designed and made for Parkinson's disease patients is remote, efficient and easy to operate.

## References

1. Su Zhi, Xie Qi, Jiang Chaowei, Zhang Fufeng. Research on Gamification of Virtual Reality Assisted rehabilitation training for elderly patients with Parkinson's disease [J]. *Hunan packaging*, 2022,37(01):121–125. DOI:<https://doi.org/10.19686/j.cnki.issn1671-4997.2022.01.031> (in Chinese).
2. Yu Tao, Kinect application development practice: talk to machines in the most natural way [M]. Beijing: China Machine Press, (2012).
3. Greg Borenstein. Making Things See; 3D vision with Kinect, Processing, Arduino, and MakerBot[M]. Maker Media, Inc, (2012).
4. J. Webb, J.Ashley. Beginning Kinect Programming with Microsoft Kinect SDK[M]. New York: Apress, (2012).
5. Jim Giles, Inside the race to hack the Kinect, *New Scientist*, Volume 208, Issue 2789, (2010), Pages 22–23.
6. Guo X, Dai Y. Occluded Joints Recovery in 3D Human Pose Estimation based on Distance Matrix[C]//2018 24th International Conference on Pattern Recognition (ICPR). IEEE, (2018): 1325–1330.
7. Liu Jianbin, Huang Minmin, Liu Zhifeng. Research on the application of Kinect based three-dimensional action reconstruction in traditional Wushu teaching and training [J]. *Sports science and technology literature bulletin*, (2022),30(02):181–182+201. DOI:<https://doi.org/10.19379/j.cnki.issn.1005-0256.2022.02.052>.
8. Zhao Wei, Li Yi. Human pose estimation optimization and animation generation based on Kinect [J / OL]. *Computer application*: 1–9 [2022–02–19]. <http://202.202.43.73:8000/rwt/CNKI/http/NNYHGLUDN3WXTLUPMW4A/kcms/detail/51.1307.TP.20211223.1351.006.html>.
9. Qian Tao Application of Kinect based dynamic posture recognition method in medical rehabilitation [D] Zhejiang: Zhejiang University of technology, 2020.
10. Su Pei, Du Weifang, & Yang Wei (2018). Research on the construction of train crew emergency training system based on VR technology *Science and technology horizon* (5), 2.
11. Wang Ligen Virtual environment system development applied to upper limb rehabilitation robot [D] Huazhong University of science and technology, 2009.
12. Qiang Lu, Li Pengfei, Su Zebin, & Jing Junfeng (2018). Customized shoe design based on VR and Kinect Computer measurement and control, 26 (4).
13. Xia min, Jiang Yijing, Zheng Dezhong, Wang Huixing, Zhan zengtu, Lin Zhicheng. Effect s of somatosensory games on cognition and gait of patients with Parkinson's disease [J]. *Clinical meta-analysis*, 2020, 35 (10): 900–903.



# Force-Directed Graph Layout Based on Community Discovery and Clustering Optimization



Linshan Han, Beilei Wang, and Songyao Wang

**Abstract** In order to visualize the important information in the knowledge graph and visualize the graph data constituting the knowledge graph for visual analysis, this paper optimizes and combines the Louvain algorithm and the force-directed graph algorithm to propose a force-directed graph layout based on community discovery and clustering optimization for the graph data. This paper uses the pruning idea to optimize the calculation steps and the community merging in the Louvain algorithm and obtains a community discovery algorithm that is more efficient and more conducive to optimizing the effect of graph layout, and introduces group elements into the force-directed graph layout to represent the community structure in graph data and implement group-based clustering optimization, so that the force-directed graph layout can clearly display the discovered community structure analyzed by the community discovery algorithm when displaying graph data, and optimize the effect and readability of the graph layout for visual analysis.

**Keywords** Visualization · Force-directed graph layout · Louvain algorithm

## 1 Introduction

The knowledge graph is an effective method for organizing knowledge in the fields of biological networks, and the visualization has also become an important technology for displaying and analyzing the information contained in the knowledge graph [1]. Data visualization technology can convert data into graphics or images to visually display the effective and valuable information in the data, which will play an important role in data analysis and mining [2]. Graph data is an important data structure in the knowledge graph. The algorithm for realizing the visualization of graph data

---

L. Han · B. Wang (✉) · S. Wang  
Northeastern University, Shenyang, Liaoning, China  
e-mail: [wangbl@swc.neu.edu.cn](mailto:wangbl@swc.neu.edu.cn)

L. Han  
e-mail: [1971115@stu.neu.edu.cn](mailto:1971115@stu.neu.edu.cn)

is the graph layout algorithm [3]. The node-link graph layout is a commonly used graph layout algorithm to reflect the data entities and entity relationships in graph data and to support graph-based network search, which uses nodes and links between nodes to reflect data entities and entity relationships [4]. When the node-link graph adopts a random layout, although the nodes and the edges of the link nodes can be drawn, the chaotic layout effect presented makes it difficult for the observer to read the information and structure in the graph data, and to make correct judgments and analysis. Therefore, in order to generate a layout with clear visual effects and easy to read and understand, common implementations of node-link graphs include force-directed graphs [5, 6], tree graphs [7, 8] and combinations with other visualization methods [9, 10].

In order to realize a graph layout that can fully mine and display the graph data constituting the knowledge graph and make the graph layout have a good layout effect [11], this paper proposes a force-directed graph layout algorithm based on community discovery and clustering optimization. Firstly, this paper uses the pruning idea to realize the community discovery process of graph compression for leaf nodes and selective community merging, thereby improving the computational efficiency and optimizing the community discovery results of the traditional Louvain algorithm. Secondly, using the clustering optimization idea, this paper introduces group elements into the force-directed graph of the traditional FR (Fruchterman-Reingold) model and implements a force-directed graph layout based on group-based clustering optimization, which optimizes the effect and readability of the graph layout.

## 2 Related Work

Community discovery is an indispensable technology for the study of the knowledge graph [12]. Louvain algorithm is a community discovery algorithm that performs well in both computational efficiency and discovery effect. In the community structure with large modularity, the similarity of nodes inside the community is high, while the similarity of nodes outside the community is low, which is a better community discovery result. Therefore, Louvain algorithm discovers a good community structure by discovering the maximum modularity, that is, by calculating the change of the modularity ( $\Delta Q$ ) and finding the maximum modularity change ( $\max \Delta Q$ ), and gradually make the community discovery result close to the community structure that maximizes modularity [13]. The Louvain algorithm can be divided into 2 stages [14]. The first stage of Louvain's algorithm regards each node in the graph data as an independent community and traverses the nodes in the graph data, and then assigns the traversed nodes to the community where each neighbor node is located and calculates  $\Delta Q$  respectively. Find  $\max \Delta Q$ , if  $\max \Delta Q > 0$ , it proves that the allocation method when  $\max \Delta Q$  is realized can increase  $Q$  and can increase  $Q$  to the maximum, so implement this allocation method and update the community structure, otherwise the community to which the node belongs will remain unchanged. When all the nodes in the graph are traversed and the communities to which all the

nodes belong no longer change, the communities are compressed into equivalent nodes. The sum of the internal edge weights of the community is converted into the self-loop edge weights of the new nodes, and the sum of the edge weights between the communities is converted into the edge weights between the corresponding new nodes, thus forming a new compressed graph. The second stage of the Louvain algorithm will continue to iterate the first stage algorithm of the Louvain algorithm until the modularity no longer increases so as to continuously approach and finally obtain the community discovery that maximizes modularity.

The force-directed graph layout is a common and popular node-link graph layout method, and its mechanical model principle can be summarized as follows: Let there be a repulsive force between any two nodes so that they are not too close together, and let there be an attractive force between two nodes with edges so that they are not too far apart. All nodes move under the interaction force, and the optimal layout is obtained when the system reaches force equilibrium and is stationary. The maturity of force-directed graph layout technology stems from the continuous research and improvement of scholars, such as the Eades model [15], the KK (Kamada-Kawai) model [16] and the FR (Fruchterman-Reingold) model [17]. In order to improve the convergence speed of the force-directed graph and optimize the final layout effect, researchers have proposed a variety of optimization strategies, such as the Multidimensional Scaling algorithm (MDS) [18], the Multilevel Algorithm [19], the Constrained Graph Layout algorithm [20], and the community gravity directed algorithm [21].

### 3 The Algorithm

#### 3.1 *Problems of Louvain Algorithm and the Optimization Idea Based on Pruning Idea*

In the traditional Louvain algorithm, it is necessary to traverse each node to calculate  $\Delta Q$ , and in the second stage, multiple iterations are required until the modularity  $Q$  does not increase anymore. Therefore, the operational efficiency of Louvain algorithm will be greatly affected when the scale of graph data is large [22]. In addition, community discovery that maximizes modularity is not exactly equivalent to community discovery that has the best visual layout, it may cause a large number of nodes in the graph data to gather in the same community after the iteration, which reduces the aesthetics of the graph layout and increases the difficulty of correct visual analysis. Therefore, in order to improve the computational efficiency of Louvain algorithm and the readability of force-oriented layout, this paper uses the optimization idea of pruning proposes an improved Louvain algorithm through graph compression for leaf nodes and selective community merging.

Graph compression for leaf nodes is a key improvement over the traditional Louvain algorithm. Define the node with only one edge in the graph data as a leaf

node. Since a leaf node can only belong to the community of its uniquely linked neighbor node, the optimization idea of pruning can be adopted to directly assign the leaf node to the community of its neighbor node, which avoids unnecessary modularity-related computations of leaf nodes and improves algorithm efficiency. Selective community merging is also a key improvement over the traditional Louvain algorithm. Since the community structure in the first stage can be regarded as a prototype of a community structure that is likely to maximize modularity, the optimization idea of pruning can be used to propose a selective community merging algorithm to replace the iterative community merging process in the second stage of the traditional Louvain algorithm, that is, a community merger method that tries to avoid mergers between large communities and promote mergers between small communities and large communities or mergers between small communities. This avoids the calculation related to the community merging method that will affect the visual layout effect, and the large amount of calculation generated by the iterative process, and improves the efficiency of the algorithm.

### 3.2 *Implementation of Improved Louvain Algorithm*

Graph data for community discovery can usually be defined as  $G(V, E)$ , where  $V$  represents a node in the data model, and  $E$  represents an edge in the data model, which is represented from a source node to a target node, and the default edge weight is 1. The implementation steps of the improved Louvain algorithm on the graph data ( $G$ ) are as follows.

Before the first stage of the improved Louvain algorithm, the graph compression for leaf nodes is performed by directly dividing the leaf nodes into the community represented by their only connected neighbor nodes, and use the method of generating equivalent nodes in the traditional Louvain algorithm to compress all nodes in each community into an equivalent node to obtain a new compressed graph. Then the same computation as the first stage of the traditional Louvain algorithm is performed on the new graph to find the community structure that maximizes modularity and compress each community into equivalent nodes.

The second stage of the improved Louvain algorithm uses selective community merging to replace the second-stage iteration process of the traditional Louvain algorithm. First, the nodes in the new graph formed by the first stage of the improved Louvain algorithm are divided into core nodes and non-core nodes. The core node is the node representing the prototype of a large community. The degree of a node can be regarded as the number of associations with other communities, the greater the degree of a node, the easier it is to become the core of a large community containing many closely related nodes. Therefore, whether a node is a core node is judged by the degree of the node, and the judgment formula is as Formula 1, where  $v$  represents the traversed node in  $G$ ,  $\text{deg}(v)$  is the weighted degree of  $v$  (equivalent to the sum of the weights of various edges of  $v$ ),  $g$  is the average of node degrees in  $G$ ,  $p$  is the standard deviation of node degrees in  $G$ .

$$\text{deg}(v) > g + p \tag{1}$$

Then traverse the non-core nodes in the new graph. For the traversed node  $i$ , in order to be more likely to achieve a community structure that maximizes the modularity, first divide the node  $i$  into the community of the neighbor core node, and find the community division that achieves  $\max \Delta Q$ . When the neighbor core node that increases  $Q$  cannot be found, in order to reduce the occurrence of too many scattered small communities, then divide the node  $i$  into the community of the neighbor non-core node, and find the community division that achieves  $\max \Delta Q$ . If the neighbor non-core node that increases  $Q$  still cannot be found, the community to which node  $i$  belongs remains unchanged. When all non-core nodes are traversed in this way and the community to which they belong does not change, a community structure with large modularity will be found.

The specific steps of the improved Louvain algorithm are as follows:

Input: Graph data  $G(V, E)$  with adjacency list structure

Output: Community structure  $C$  of graph data  $G$

Step 1: Initialize each node in  $G$  as a community and traverse the nodes in  $G$ .

Step 2: For node  $i$  ( $i \in V$ ), when  $i$  is a leaf node, divide  $i$  directly into its neighbor community; when  $i$  is a non-leaf node, if  $\Delta Q > 0$  exists after  $i$  is divided into its neighbor community, divide  $i$  into the community that realizes  $\max \Delta Q$ ; otherwise, the community of  $i$  remains unchanged.

Step 3: Repeat Step 2 until the community structure no longer changes.

Step 4: Save the community structure  $C$  at this time, compress each community into each equivalent point to obtain a compressed new graph  $G'(V', E')$ , traverse the nodes in  $G'$ , determine whether the node is a core node, and store it in the core node set  $K$  and the non-core node set  $N$  ( $N = V' - K$ ).

Step 5: Traverse the nodes in the non-core node set  $N$ . For node  $i'$  ( $i' \in N$ ), when  $i'$  is a leaf node, divide  $i'$  into its neighbor community directly; when  $i'$  is a non-leaf node, if  $\Delta Q > 0$  exists after  $i$  is divided into its neighbor core community, divide  $i$  into the community that realizes  $\max \Delta Q$ ; otherwise, if  $\Delta Q > 0$  exists after  $i$  is divided into its neighbor non-core community, divide  $i$  into the community that realizes  $\max \Delta Q$ ; otherwise, the community of  $i$  remains unchanged.

Step 6: Repeat Step 5 until the community structure no longer changes.

Step 7: Update and output the community structure  $C$  at this time. The algorithm is over.

### ***3.3 Problems of Force-Directed Graph Layout and the Clustering Optimization Idea***

The existing optimization strategies of force-directed graph layout usually only adjust the mechanical model, and the geometric distance of nodes in the layout formed by the mechanical model often has a certain error with the path length between nodes in

the graph data. Moreover, when optimizing the mechanical model, it often only relies on the basic attributes of nodes and edges in the graph data, and does not analyze and mine the information contained in the graph data. These will not only make the final graph layout insufficient to display the graph data analysis results, but also affect the understanding and judgment of observers when using the graph layout for visual analysis.

Since the community structure is an important data analysis attribute of the knowledge graph and the division of the community of nodes in the graph data by the community discovery algorithm is very consistent with the node clustering of the force-oriented graph layout, this paper introduces group elements into the FR model to realize the optimization and adjustment of the mechanical model, and let the grouping elements correspond to the community structure obtained by the improved Louvain algorithm. This optimizes the layout effect and the optimized direction is more consistent and fully displays the information contained in the graph data.

### 3.4 Group-Based Clustering Optimization Implementation

In order to enable the force-directed graph layout to display the community structure of graph data, this paper adds group elements representing the community on the basis of the node elements and edge elements contained in the force-directed graph layout of the traditional FR model to guide the clustering optimization of the force-directed graph layout, which is implemented as follows:

The grouping element represents the community entity in the graph data. Just like the Louvain algorithm can regard the community as an equivalent point, the model and rendering method of group elements can also be implemented with reference to node elements in the force-directed graph layout, which is represented as a circle surrounding the node element in the group. In order to achieve group-based clustering optimization, it is necessary to make the nodes in the same group as clustered as possible, and to avoid overlapping between different groups as much as possible. Therefore, this paper defines the force-directed model based on the FR model.

There is a repulsive force  $F_r$  between all nodes, and an attractive force  $F_a$  between nodes connected by edges.

$$\begin{cases} F_r = \frac{K_r}{r^2} \\ F_a = \frac{K_a \times r}{\partial(c)} \end{cases} \quad (2)$$

The calculation formula is as Formula 2, where  $r$  is the distance between two nodes;  $K_r$  and  $K_a$  are the coefficients that control the strength of the attractive and repulsive forces; in order to distinguish different communities, it is necessary to keep the nodes of different communities as far away as possible, that is, to attenuate the attractive force between nodes that are connected by edges but belong to different

communities, so  $\partial(c)$  indicates whether the nodes belong to the same community, the same is 1, otherwise it is the attenuation coefficient  $C$ .

In order to make the nodes in the same group gather as much as possible, this paper refers to the force-directed model with gravity, and adds an intra-group gravity  $F_c$  to the grouped element from the average position center of the nodes in the group.

$$F_c = G \times \text{deg}(v) \tag{3}$$

The calculation formula is as Formula 3, where  $G$  is the gravity coefficient that controls the tightness of the internal layout of the group;  $\text{deg}(v)$  is the degree of the node, which represents the core degree of the node, making the core node in the community closer to the center of the community element.

## 4 Experimental Results and Analysis

In order to verify the algorithm implemented in this paper, this paper uses real large-scale undirected graph datasets such as Zachary karate club, Dolphins, Lesmis and Facebook as the validation of this experiment. The characteristics of these datasets are shown in Table 1. We experimented and compared the community discovery effects of the traditional Louvain algorithm and the improved Louvain algorithm proposed in this paper on these datasets. The evaluation indicators of the effect are the modularity  $Q$ , the number of communities  $N$ , and the running time  $T$  of the algorithm.  $Q$  and  $N$  reflect the discovered community structure. The larger the  $Q$ , the better the obtained community structure, and the smaller the  $T$ , the higher the efficiency of the algorithm.

The experimental results are shown in Table 2. It can be seen that the modularity  $Q$  obtained by the improved Louvain algorithm is improved compared with the traditional Louvain algorithm, which means that a better community structure can be found. And the computational efficiency of the improved algorithm is improved, and as the scale of the graph dataset becomes larger, the computational efficiency is improved more significantly.

Then we visualized the Zachary karate club, Dolphins and Lesmis datasets and used colors to distinguish different communities of the graph data. We adopted three forms of graph layouts for visualization, namely the traditional louvain algorithm,

**Table 1** Dataset information

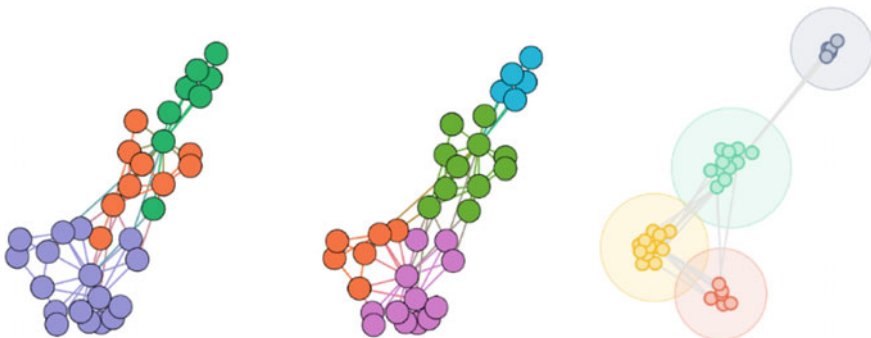
Dataset information	Dataset name			
	Zachary karate club	Dolphins	Lesmis	Facebook
Number of nodes	34	62	77	2888
Number of edges	78	159	254	2981
Proportion of leaf nodes (%)	2.94	14.52	22.08	96.61

**Table 2** The experimental results

Evaluation index	Dataset	Improved Louvain	Traditional Louvain
$Q$	Zachary karate club	0.4198	0.3807
	Dolphins	0.5188	0.4955
	Lesmis	0.5583	0.5006
	Facebook	0.8087	0.8086
$N$	Zachary karate club	4	3
	Dolphins	5	4
	Lesmis	6	5
	Facebook	8	8
$T$ (s)	Zachary karate club	0.001990	0.002950
	Dolphins	0.003751	0.005941
	Lesmis	0.004032	0.009932
	Facebook	0.064841	16.556743

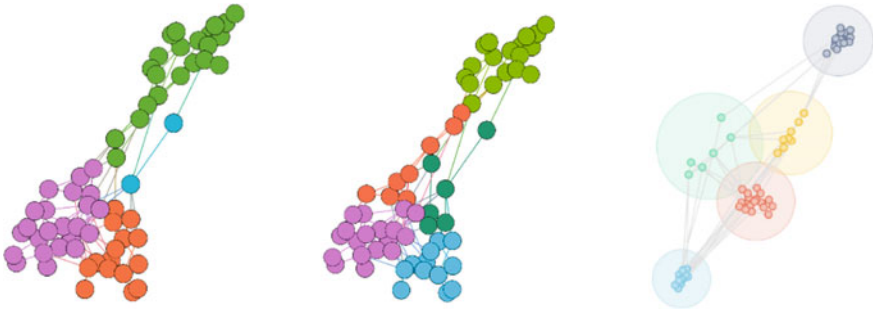
(b) the traditional force-directed graph layout for the improved louvain algorithm, and (c) the force-directed graph layout based on group-based clustering optimization for the improved Louvain algorithm. The visualization results are shown in Figs. 1, 2 and 3 in turn.

As can be seen from Figs. 1, 2 and 3, the community structure found by the improved Louvain algorithm is obvious, and the distribution of nodes among the communities is more uniform. And the improved grouping-based clustering optimization force-directed graph layout can further highlight the community structure of graph data, which is helpful for observers to perform visual analysis.



**Fig. 1** The three forms of graph layouts for the Zachary karate club





**Fig. 2** The three forms of graph layouts for the Dolphins



**Fig. 3** The three forms of graph layouts for the Zachary karate club

## 5 Conclusion and Future Work

This paper optimizes and combines the Louvain algorithm of the traditional community discovery algorithm and the force-directed graph of the classical node-link graph, and proposes a force-directed graph layout based on community discovery and clustering optimization. Experiments show that the method in this paper has a higher rate of community discovery, the discovered community structure has better modularity and visualization effects, clustering optimization guided by the obtained community structure can improve the readability of force-directed graph layout, and while optimizing the layout effect, it is more helpful to analyze and understand the graph data. In the future, we will continue to research and improve this algorithm, such as further optimizing and reducing the influence of overlapping graphics on the effect of graph layout and establishing a more comprehensive and objective visual layout evaluation index to guide the realization of automatic map layout.

## References

1. Yan J, Wang C, Cheng W, et al.: A retrospective of knowledge graphs. *Frontiers of Computer Science* 12(1), (2018).
2. Ren L, Du Y, Ma S, et al.: Visual analytics towards big data. *Ruan Jian Xue Bao/Journal of Software* 25(9), 1909–1936 (2014).
3. Wang Yongchao, Luo Shengwen, Yang Yingbao, et al.: A Survey on Knowledge Graph Visualization. *Journal of Computer-Aided Design & Computer Graphics* 31(10), 1666–1676 (2019).
4. Liu, S., Xiao, Z., You, X. and Su, R., 2022. Multistrategy boosted multicolony whale virtual parallel optimization approaches. *Knowledge-Based Systems*, 242, p. 108341.
5. Wang Y, Wang Y, Sun Y, et al.: Revisiting Stress Majorization as a Unified Framework for Interactive Constrained Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics* 24(1), 489–499 (2018).
6. Ashley Suh, Mustafa Hajij, Bei Wang, et al.: Persistent Homology Guided Force-Directed Graph Layouts. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS* 26(1), 697–707 (2020).
7. Jochen Gortler, Christoph Schulz, Daniel Weiskopf, et al.: Bubble Treemaps for Uncertainty Visualization. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS* 24(1), 719–728 (2018).
8. Ge H.A, Yong L.B, Xu T.C, et al.: PLANET: A radial layout algorithm for network visualization. *Physica A* 539, (2020).
9. Holger Stitz, Samuel Gratzl, Harald Piringer, et al.: KnowledgePearls: Provenance-Based Visualization Retrieval. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS* 25(1) 120–130 (2019).
10. Timothy M, Basole R C: Graphicle: Exploring Units, Networks, and Context in a Blended Visualization Approach. *IEEE Transactions on Visualization and Computer Graphics* 25(1), 576–585 (2019).
11. Su R., Gu, Q. and Wen, T., 2014. Optimization of high-speed train control strategy for traction energy saving using an improved genetic algorithm. *Journal of Applied Mathematics*, 2014.
12. Rieck B, Fugacci U, Lukaszczuk J, et al.: Clique Community Persistence: A Topological Visual Analysis Approach for Complex Networks. *IEEE Transactions on Visualization & Computer Graphics* 24(1), 822–831 (2018).
13. VINCENT D B, GUILLAUME J L, RENAUD L, et al.: Fast unfolding of communities in large network. *Journal of Statistical Mechanics: Theory and Experiment* 10, 1–12 (2008).
14. WU Zu-feng, WANG Peng-fei, QIN Zhi-guang, et al.: Improved Algorithm of Louvain Communities Dipartition. *Journal of University of Electronic Science and Technology of China* 42(1), 105–108 (2013).
15. EADES P: A heuristic for graph drawing. *Congressus numerantium* 42, 149–160 (1984).
16. KAMADA T, KAWAI S, et al.: An algorithm for drawing general undirected graphs. *Information processing letters* 31(1), 7–15 (1989).
17. FRUCHTERMAN T M J, REINGOLD E M: Graph drawing by force-directed placement. *Software Practice & Experience* 21(1) 1129–1164 (1991).
18. Khoury M, Hu Y, Krishnan S, et al.: Drawing Large Graphs by Low-Rank Stress Majorization. *Computer Graphics Forum*, (2012).
19. Yoghoudjian V, Dwyer T, Klein K, et al.: Graph Thumbnails: Identifying and Comparing Multiple Graphs at a Glance. *IEEE Transactions on Visualization and Computer Graphics* 24(12), 3081–3095 (2018).
20. Yunhai Wang, Mingliang Xue, Yanyan, et al.: Wang Interactive Structure-aware Blending of Diverse Edge Bundling Visualizations. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS* 26(1), 687–696 (2020).
21. Wu Yu, Li Zaoxu, Li Hongbo, et al.: A community-gravity directed algorithm for showing community structure of complex networks. *Journal of Computer-Aided Design & Computer Graphics* 27(8), 1460–1467 (2015).

22. Hao Runqian, Wu Yu, Chen Xin: An Algorithm for Large-scale Social Network Community Detection and Visualization. *Journal of Computer-Aided Design & Computer Graphics* 29(2), (2017).

# Comprehensive Strategy to Screen the Ankylosing Spondylitis-Related Biomarkers in the Peripheral Serum



Zhenrun Zhan, Xiaodan Bi, Xu Tang, and Tingting Zhao

**Abstract** Ankylosing spondylitis (AS) is one of the main research priorities in spine surgery and rheumatology, and available research suggests that AS can be strongly associated with genetic, environmental, immunological and endocrine pathological factors. It is meaningful that exploring differential genes associated with ankylosing spondylitis (AS) through bioinformatics strategies to find novel diagnostic markers for the disease. **Methods:** The GSE25101 dataset was downloaded from the GEO, screened for DEGs, and then subjected to GO and KEGG enrichment analysis. Next, five algorithms, WGCNA, RF machine learning algorithm, SVM-RFE algorithm, protein–protein interaction network (PPI) analysis, and LASSO logistic regression were used to screen new and critical biomarkers of AS. **Results:** We performed GO, KEGG, DO and GSEA enrichment analysis with 62 screened DEGs to explore their interactions with biological functions, mechanisms of action and related diseases. The results suggest that multiple signaling pathways enriched by DEGs may be intimately involved in the onset and procession of ankylosing spondylitis. In addition, combining multiple algorithms, PTPN1 was determined to be a promising biomarker in the serum of AS patients and showed good diagnostic value. **Conclusion:** In conclusion, we used a holistic approach to select for biomarker associated with ankylosing spondylitis, and PTPN1 may serve as a novel diagnostic marker associated with peripheral blood AS disease.

**Keywords** Ankylosing spondylitis (AS) · Machine learning algorithm · Biomarker · Differentially expressed genes · Weighted gene co-expression network analysis (WGCNA)

---

Z. Zhan · X. Bi · X. Tang · T. Zhao (✉)

Heping Hospital Affiliated to Changzhi Medical College, Changzhi, Shanxi, China

e-mail: [649823325@qq.com](mailto:649823325@qq.com)

Changzhi Medical College, Changzhi, Shanxi, China

## 1 Introduction

Ankylosing spondylitis (AS) is a disabling disease of unknown etiology with inflammation of the sacroiliac joints and spinal attachment points as the main symptom, mostly seen in men, and its onset tends to be younger [1]. AS is one of the research priorities in spine surgery and rheumatology, and the available studies have shown that AS is mainly associated with genetic, environmental, immunological and endocrine pathological factors [2–4]. Currently, biomarkers for the early diagnosis and treatment of AS are still lacking in the clinical setting, and therefore it is necessary to explore important targets related to AS. As technology continues to mature and evolve, whole genome sequencing for many diseases is available, in recent year. One study used cRNA microarray technology for detecting cRNA expression in peripheral blood of AS patients and found that ring RNA may be an important marker for the diagnosis and progression observation of AS [5]. Application of single-cell sequencing technology in AS disease further reveals the underlying mechanisms of AS progression [6]. Bioinformatics analysis can be used to screen for promising biomarkers for various oncological and non-oncological diseases [7, 8]. First, we downloaded the GSE25101 dataset from NCBI's GEO database. Then, we screened differentially expressed genes from it, further constructed a weighted gene co-expression network, and combined with other algorithms to select core genes, which are the candidate markers we were looking for. Finally, we explored the possible functions of the biomarkers through various functional and pathway enrichment analyses to offer novel ideas on early diagnostics and therapies for AS.

## 2 Materials and Methods

### 2.1 Data Collection and Processing

First, the AS-related datasets were searched in the database and then downloaded by applying the GEOquery [9] package, and finally the GSE25101 dataset containing the AS group and the normal control group was downloaded. Next, we annotated the data and finally obtained the complete dataset information for the next step of analysis.

### 2.2 DEGs Screening and GSEA Analysis

First, we first performed differential expression analysis of genes, and our screening conditions were set to  $P < 0.05$  and  $|\log_2(\text{FC})| > 0.5$ , and those satisfying this screening condition were identified as DEGs. Next, the heat map was depicted by

the “pheatmap” package [10]. Meanwhile, the GSEA analysis of the screened target gene set was performed by applying R software.

### ***2.3 WGCNA-Based Targeting Modules and Genes Screening***

First, a weighted gene co-expression network was constructed using the “WGCNA” package [11]. Set mean FPKM = 0.5 as the filtering criterion, load the trait data, and evaluate the similarity of gene expression at the same time to obtain the matrix file. We set soft thresholds, minimal gene modules, visualized the gene network graph, and finally correlated the modules with clinical traits to filter out the target modules and genes closely related to AS.

### ***2.4 Functional Enrichment Analysis***

In R software, the clusterprofiler package is loaded and run it for GO, KEGG, and DO analysis of the target genes [12].

### ***2.5 PPI Network Analysis***

PPI refers to the non-covalent binding of two or more protein molecules. The PPI network of candidate DEGs was analyzed using STRING software and the confidence score was set to 0.4. In addition, cytoscape was chosen to validate the net and screen for top ten genes.

### ***2.6 Multiple Algorithms Combine to Identify Prospective Biomarkers of AS***

Five algorithms were used to screen new and critical biomarkers for ankylosing spondylitis, including random forest (RF) machine learning algorithms [13], Protein–Protein Interaction Networks (PPI) analysis, LASSO logistic regression [14], SVM-RFE [15], and WGCNA. The Random Forest package from R software has been employed to establish a random forest model based on candidate genes, the average misjudgment rate of all the genes was calculated on the basis of the out-of-band data. Then, the glmnet package was applied to construct the LASSO model to filter the candidate mRNAs from the dataset [12]. The ROC curves were constructed, and the area under the curve was applied to reflect the predictive efficacy of the model. In the

next step, the “e1071” package [11] was loaded in the R software and we employ it to filter suitable genes using the SVM-RFE algorithm. Lastly, the core genes selected by the integrated strategy are recognized as biomarkers for AS.

## ***2.7 Validation of the Diagnosis-Related Gene Expression and Regulatory Mechanisms and Biological Functions of the Potential Biomarkers***

In the R software, the pROC package was loaded to reflect the diagnostic efficacy of candidate genes by drawing ROC curves [16]. The differences in biomarker expression levels between the two groups were also analyzed. Next, GSEA analysis of candidate genes was carried out to explore their potential relationships with various functions and pathways. At the same time, we divided the samples into AS and normal groups and used the ssGSEA R package “GSVA” to explore the different levels of HALLMARK pathways in the expression profile of the GSE25101 dataset.

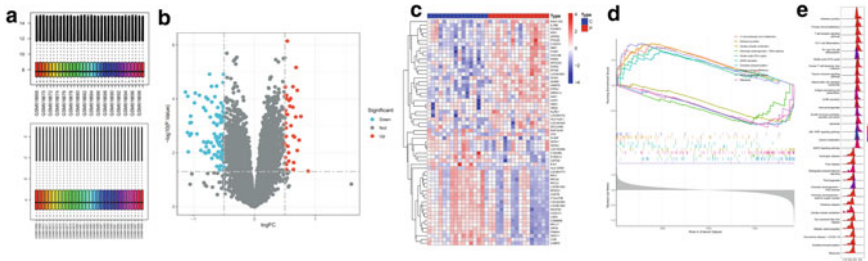
## **3 Results**

### ***3.1 Identification of DEGs of Patients with Ankylosing Spondylitis***

After gene annotation and data standardization (Fig. 1a), we used  $|\log_2 \text{FC}| > 0.5$  and adjusted  $P$  value  $< 0.05$  as inclusion conditions. We found 62 DEGs from the GSE25101 dataset, including 20 up-regulated mRNAs and 42 down-regulated mRNAs, and drew a volcano plot containing this information (Fig. 1b). In the R software, the heatmap package was loaded for drawing the differentially expressed mRNAs in the heatmap (Fig. 1c). The GSEA enrichment analysis illustrated that GnRH secretion, Citrate cycle (TCA cycle), Adherens junction, Primary immunodeficiency and 2-Oxocarboxylic acid metabolism were predominantly positive in AS samples, while Renin-angiotensin system, Oxidative phosphorylation, Chemical carcinogenesis—DNA adducts, Ribosome and Cardiac muscle contraction were highly active in normal samples (Fig. 1d, e).

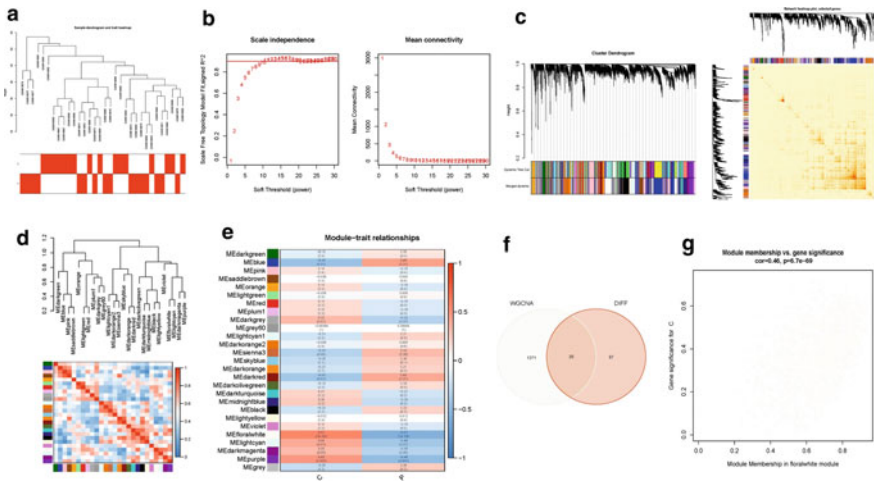
### ***3.2 WGCNA-Based Filtering of Goal Modules and mRNAs***

We incorporated 12,926 genes for the construction of a weighted gene co-expression network (Fig. 2a) and selected the most appropriate soft threshold (Fig. 2b). We merged modules with a feature factor greater than 0.7 and limited each module



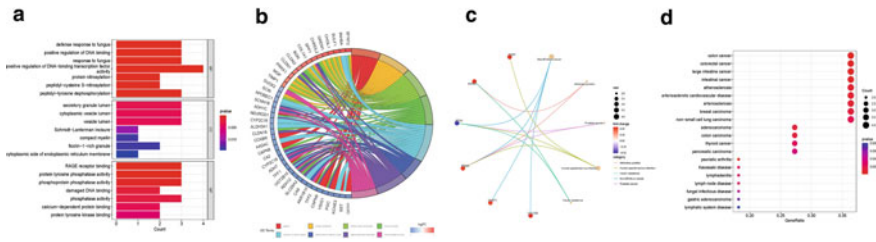
**Fig. 1** The differential gene expression of GSE25101. **a** The box graphs of mRNA expression in normal and AS samples. Volcano plot (**b**) and expression heat map (**c**) of differential mRNAs. **d** Display of the top 10 signaling pathways in the GSEA-based dataset. **e** The ridgeplot of the result on GSEA

to contain no less than 50 genes, screening out 27 modules (Fig. 2c). In addition, we investigated the correlation of the modules (Fig. 2d). Lastly, our MEfloralwhite module got chosen as the most relevant candidate module for the occurrence of AS (Fig. 2e), and 25 differentially expressed mRNAs were singled out (Fig. 2f).



**Fig. 2** Characterization of candidate mRNAs by WGCNA. **a** The dendrogram and trait heatmap of 32 samples. **b** Soft threshold analysis. **c** Target modules as determined after performing merging and filtering. **d** Relevance between the 27 modules. **e** Associations between AS and modules. **f** Venn diagram of the intersection of MEfloralwhite genes and differential genes. **g** The scatter graph reflect the association between the gene importance of AS and the FLORALWHITE module





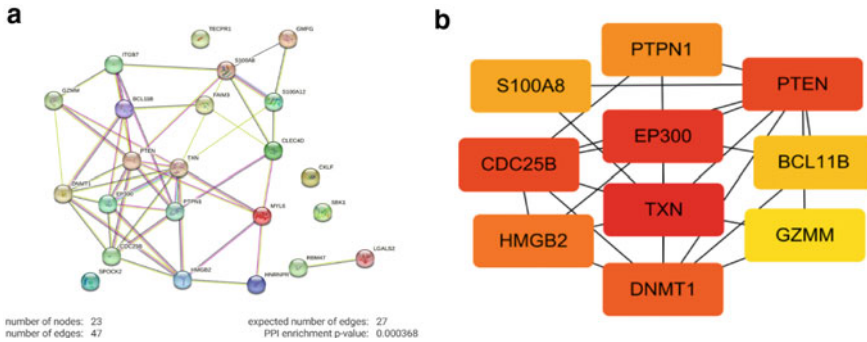
**Fig. 3** **a** GO outcomes of the candidate genes; **b** ring plot showing GO enrichment and relevant DEGs; **c** results of KEGG enrichment analysis; **d** bubble chart showing the DO enrichment results

### 3.3 GO, KEGG and DO Enrichment Analyses

The clusterProfiler package was loaded into the R software and 25 genes were analyzed for GO, DO and KEGG enrichment for bio-functional studies. The analysis results of GO enrichment were shown in Fig. 3a. Positive regulation of protein kinase activity, positive regulation of response to external stimulus, positive regulation of kinase activity, regulation of DNA-binding transcription factor activity and Leukocyte migration were associated with biological processes (BP). Cytoplasmic vesicle lumen, vesicle lumen and secretory granule lumen were involved in the cellular components (CC). In terms of molecular function (MF), genes were enriched in phosphatase activity, phosphate hydrolase activity and receptor ligand activity signaling receptor activator activity. Figure 3b illustrates a considerable fraction of the GO-enriched words and associated DEGs. Furthermore, the DEGs were mainly concentrated in MicroRNAs in cancer, Adherens junction, Prostate cancer and Human papillomavirus infection in the KEGG enrichment analysis (Fig. 3c). Moreover, the DEGs were primarily clustered in colon cancer, arteriosclerotic cardiovascular disease, arteriosclerosis and breast carcinoma in the DO-rich concentration (Fig. 3d).

### 3.4 Establishing a PPI Network

We opened STRING, entered the target genes, and analyzed the interactions for the proteins they translated. We can see that the constructed PPI network has 23 nodes and 47 edges (Fig. 4a). Moreover, we selected 10 core genes that are most closely related to each other (Fig. 4b). In particular, the PPI network is built and visualized by the Cytoscape v3.9.0 software.



**Fig. 4** **a** PPI network of the candidate mRNAs; **b** ten core genes of PPI network

### 3.5 Machine Learning Algorithm-Based Recognition of Prospective Biomarkers for AS

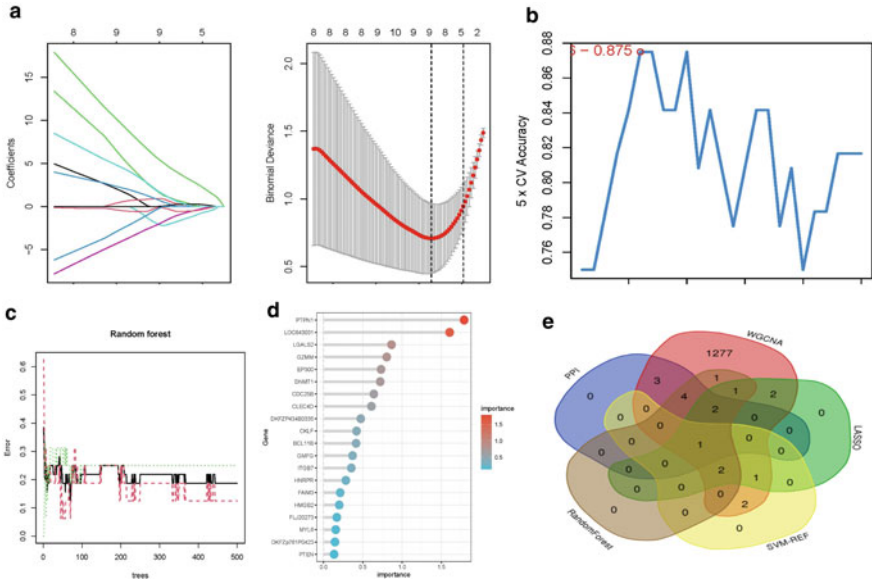
We identified 9 genes as biomarkers for AS from DEGs using the LASSO logistic regression algorithm (Fig. 5a). Six mRNAs were detected as diagnostic markers from the DEGs with the SVM-RFE algorithm (Fig. 5b). 11 mRNAs have been defined as significant biomarkers for the RF algorithm (Fig. 5c, d). To identify highly correlated sets of genes in their expression blocks, we conducted a joint analysis using data obtained by the PPI and WGCNA algorithms. At last, we obtained PTPN1 obviously correlated to AS through the overlap of five algorithms (Fig. 5e).

### 3.6 Validation of the Diagnosis-Related Gene Expression

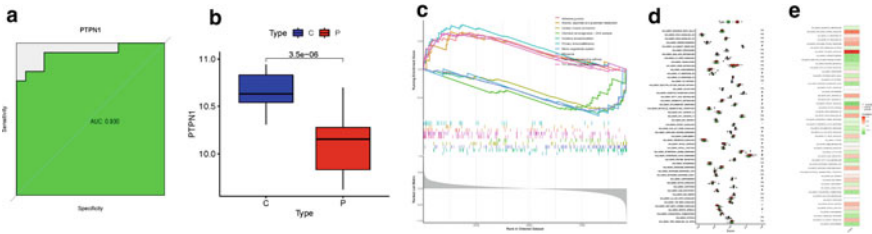
To validate the potential of PTPN1 as diagnosis marker for AS, we performed ROC analysis of this gene in the expression dataset GSE25101 and plotted ROC curves (Fig. 6a). At the same time, to further validate the role of PTPN1 in AS, we also looked at the expression of candidate mRNAs in people with ankylosing spondylitis. Intriguingly, we derived an observation that PTPN1 expression is inhibited in AS patients (Fig. 6b).

### 3.7 Regulatory Mechanisms and Biological Functions of the Potential Biomarkers

We conducted GSEA study on the ordered gene expression matrix to explore the regulatory effect of PTPN1. The analysis results show that PTPN1 was mainly involved in B cell receptor signaling pathway, Complement and coagulation



**Fig. 5** Screening for hub biomarkers through a comprehensive strategy: **a** the LASSO algorithm was performed to reserve the highest promising genes; **b** results of candidate mRNAs by SVM-RFE; **c** Filtering mRNAs by random forest (RF) algorithm; **d** results of candidate mRNAs by RF; **e** the image displays the convergence of the diagnostic markers acquired via five algorithm



**Fig. 6** **a** ROC curves of PTPN1; **b** the expression levels of PTPN1 in AS patients and control samples; ssGSEA analysis: **c** analysis results of PTPN1; **d** expression of genes of different HALLMARK pathways in AS and normal samples; **e** correlation analysis of PTPN1 and pathways

cascades, Primary immunodeficiency, Chemical carcinogenesis-DNA adducts and SNARE interactions in vesicular transport (Fig. 6c). Then, we employed ssGSEA to analyze the data of 32 samples obtained from GSE25101 to select their enriched HALLMARK pathways, then the richness levels of 48 HALLMARK pathways in AS and normal samples were obtained. For further study, we performed computational matrix scores, comparative differences, and single-gene correlation tests. As shown in Fig. 6d, e, the expression of gene, which up-regulated by ROS and activation of WNT signaling through accumulation of beta catenin CTNNB1, up-regulated

through activation of mTORC1 complex, involved in cholesterol homeostasis, in response to TGF $\beta$ 1 was statistically different between AS and normal samples. The most relevant pathways to PTPN1 are HALLMARK\_TGF\_BETA\_SIGNALING, HALLMARK\_REACTIVE\_OXYGEN\_SPECIES\_PATHWAY and HALLMARK\_CHOLESTEROL\_HOMEOSTASIS.

## 4 Discussion

AS is a common clinical condition and has a serious impact on the daily life of patients, and in severe cases, it can even lead to disability. Therefore, it is important to investigate the mechanisms of AS by biological information technology and to explore biomarkers associated with AS. For this research, bioinformatics was employed to identify and evaluate the gene expression of healthy individuals and AS patients, and screened 62 differentially expressed genes. GO functional analysis revealed the DEGs were primarily engaged in biological processes including oxidative phosphorylation metabolism. Oxidative stress damage was evident in AS patients, and oxidative stress damage contributes to sickness activation [17]. It has been shown that anti-inflammatory treatment can significantly improve the activity of various lipids and enzymes, which may become one of the directions for the treatment of AS [18]. KEGG signaling pathway enrichment analysis revealed DEGs were primarily concentrated in MicroRNAs in cancer, Human papillomavirus infection and other signaling pathways. Chen-Yu Wei et al. showed through a prospective clinical study that patients infected with HPV are more likely to develop ankylosing spondylitis [19]. Mohammadi et al. [20] reviewed the recent studies on the role of miRNAs in the development of AS and explored the possibility of applying miRNAs as prognostic markers and targeting it for therapeutic strategies. For example, regulation of miR-495 can then influence the development of ankylosing spondylitis from the mechanism of programmed death [21]. PPI network analysis, ROC curve analysis, three machine learning algorithms and gene expression analysis further clarified that the gene PTPN1 can be used as a diagnostic marker for AS. The protein tyrosine phosphatase nonreceptor type 1 (PTPN1) gene encode the protein tyrosine phosphatase 1B (PTP1B). The target of PTP1B is the leptin receptor, and its signaling mediates the metabolism of glucose [22]. But the association of this gene with the pathogenesis of ankylosing spondylitis has not been adequately studied. It is suggested that further study of this pathway may be a potential therapeutic direction for AS. According to the results of GSEA analysis, PTPN1 is mainly invoked in the B-cell receptor signaling pathway and primary immunodeficiency, and its expression was mainly up-regulated by reactive oxygen species (ROS) and activation of mTORC1 complex. The results of Jennifer J. Schwarz et al. also confirm the character of this gene in the B-cell receptor signaling pathway [23]. The negative regulatory effect of PTPN1 on immunity has also been experimentally confirmed in the past [24]. Previous studies have demonstrated that in patients with ankylosing

spondylitis, there may be mitochondrial dysfunction and excess ROS causing senescence of mesenchymal cells, and these are consistent with the conclusions drawn from our raw letter analysis [25]. In summary, we have studied both the onset and progression of AS in greater depth, which points us to the next step in developing a treatment for AS.

Screening for disease-related diagnostic markers can provide direction for early diagnosis and treatment of disease. Yu et al. used chromatography mass spectrometry to screen hip ligament samples from AS and non-AS groups for differential expression of myeloperoxidase and non-AS hip ligament samples, and found that myeloperoxidase might Myeloperoxidase may be an important marker associated with AS-induced hip lesions [26]. The use of whole blood samples for screening disease-related diagnostic markers is more convenient than obtaining tissue samples. García-Salinas et al. [27] evaluated human leukocyte antigen B27 (HLA-B27) as a diagnostic marker for axial spondylitis in a cohort study, and the result showed that HLA-B27 has good specificity but low sensitivity for the diagnosis of axial spondylitis disease. In this study, we screened the peripheral blood diagnostic marker PTPN1 in AS patients by bioinformatics analysis method to further complement the existing diagnostic methods. This marker may be a potential target for treatment of AS and may be an important modality for future non-surgical treatment of AS, reducing the risk of treatment modalities such as tissue engineering repair and surgical procedures.

**Acknowledgements** This study was supported by Shanxi Province Graduate Education Innovation Project (2022Y737), the Basic Research Program of Shanxi Province (202203021212010), and the Youth Start-up Fund of Heping Hospital affiliated to Changzhi Medical College (HPYJ202225).

## References

1. Crossfield SSR, Marzo-Ortega H, Kingsbury SR, Pujades-Rodriguez M, Conaghan PG. Changes in ankylosing spondylitis incidence, prevalence and time to diagnosis over two decades. *RMD Open*. 2021;7(3):e001888. doi: <https://doi.org/10.1136/rmdopen-2021-001888>
2. Schlosstein L, Terasaki PI, Bluestone R, Pearson CM. High association of an HL-A antigen, W27, with ankylosing spondylitis. *N Engl J Med*. 1973;288(14):704–706. doi: <https://doi.org/10.1056/NEJM197304052881403>
3. Wu Y, Ren M, Yang R, et al. Reduced immunomodulation potential of bone marrow-derived mesenchymal stem cells induced CCR4+CCR6+ Th/Treg cell subset imbalance in ankylosing spondylitis. *Arthritis Res Ther*. 2011;13(1):R29. Published 2011 Feb 21. doi: <https://doi.org/10.1186/ar3257>
4. Kebapcilar L, Bilgir O, Alacacioglu A, et al. Impaired hypothalamo-pituitary-adrenal axis in patients with ankylosing spondylitis. *J Endocrinol Invest*. 2010;33(1):42–47. doi: <https://doi.org/10.1007/BF03346548>
5. Tang YP, Zhang QB, Dai F, et al. Circular RNAs in peripheral blood mononuclear cells from ankylosing spondylitis. *Chin Med J (Engl)*. 2021;134(21):2573–2582. Published 2021 Oct 19. doi: <https://doi.org/10.1097/CM9.0000000000001815>

6. Xu H, Yu H, Liu L, et al. Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Peripheral Mononuclear Cells in Patients With Ankylosing Spondylitis. *Front Immunol.* 2021;12:760381. Published 2021 Nov 22. doi: <https://doi.org/10.3389/fimmu.2021.760381>
7. Zhan, Z., Zhao, T., Bi, X., Yang, J., Han, P. (2022). Identification and Evaluation of Key Biomarkers of Acute Myocardial Infarction by Machine Learning. In: Huang, DS., Jo, KH., Jing, J., Premaratne, P., Bevilacqua, V., Hussain, A. (eds) *Intelligent Computing Theories and Application. ICIC 2022. Lecture Notes in Computer Science*, vol 13394. Springer, Cham. [https://doi.org/10.1007/978-3-031-13829-4\\_9](https://doi.org/10.1007/978-3-031-13829-4_9)
8. Wang T, Zheng X, Li R, et al. Integrated bioinformatic analysis reveals YWHAB as a novel diagnostic biomarker for idiopathic pulmonary arterial hypertension. *J Cell Physiol.* 2019;234(5):6449–6462. doi: <https://doi.org/10.1002/jcp.27381>
9. Meltzer, D.P.S.: GEOquery: A bridge between the gene expression omnibus (GEO) and BioConductor. *Bioinformatics.* 23, 1846–7 (2007).
10. Ritchie, M.E., et al.: Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47 (2015)
11. M.-L. Huang, Y.-H. Hung, W. M. Lee, R. K. Li, and B.-R. Jiang, “SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier,” *Scientific World Journal*, vol. 2014, article 795624, pp. 1–10, 2014.
12. Yu, G., Wang, L.G., Han, Y., He, Q.Y.: clusterProfiler: An r package for comparing biological themes among gene clusters. *Omics-a Journal of Integrative Biology.* 16, 284–287 (2012).
13. Alhamzawi R, Ali HTM. The Bayesian adaptive lasso regression. *Math Biosci.* 2018;303:75–82. doi: <https://doi.org/10.1016/j.mbs.2018.06.004>
14. Alakwaa FM, Chaudhary K, Garmire LX. Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. *J Proteome Res.* 2018;17(1):337–347. doi: <https://doi.org/10.1021/acs.jproteome.7b00595>
15. Lin X, Yang F, Zhou L, et al. A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *J Chromatogr B Analyt Technol Biomed Life Sci.* 2012;910:149–155. doi: <https://doi.org/10.1016/j.jchromb.2012.05.020>
16. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., Müller, M.: pROC: An open-source package for r and s+ to analyze and compare ROC curves. *BMC Bioinformatics.* 12, 77 (2011).
17. Solmaz D, Kozacı D, Sarı İ, et al. Oxidative stress and related factors in patients with ankylosing spondylitis. *Eur J Rheumatol.* 2016;3(1):20–24. doi: <https://doi.org/10.5152/eurjrheum.2015.0031>
18. Czókolyová M, Pusztaí A, Végh E, et al. Changes of Metabolic Biomarker Levels upon One-Year Anti-TNF- $\alpha$  Therapy in Rheumatoid Arthritis and Ankylosing Spondylitis: Associations with Vascular Pathophysiology. *Biomolecules.* 2021;11(10):1535. Published 2021 Oct 18. doi: <https://doi.org/10.3390/biom11101535>
19. Wei CY, Lin JY, Wang YT, Huang JY, Wei JC, Chiou JY. Risk of ankylosing spondylitis following human papillomavirus infection: A nationwide, population-based, cohort study. *J Autoimmun.* 2020;113:102482. doi: <https://doi.org/10.1016/j.jaut.2020.102482>
20. Mohammadi H, Hemmatzadeh M, Babaie F, et al. MicroRNA implications in the etiopathogenesis of ankylosing spondylitis. *J Cell Physiol.* 2018;233(8):5564–5573. doi: <https://doi.org/10.1002/jcp.26500>
21. Ni WJ, Leng XM. Down-regulated miR-495 can target programmed cell death 10 in ankylosing spondylitis. *Mol Med.* 2020;26(1):50. Published 2020 May 25. doi: <https://doi.org/10.1186/s10020-020-00157-3>
22. Bence KK, Delibegovic M, Xue B, et al. Neuronal PTP1B regulates body weight, adiposity and leptin action [published correction appears in *Nat Med.* 2010 Feb;16(2):237]. *Nat Med.* 2006;12(8):917–924. doi: <https://doi.org/10.1038/nm1435>
23. Schwarz JJ, Grundmann L, Kokot T, et al. Quantitative proteomics identifies PTP1B as modulator of B cell antigen receptor signaling. *Life Sci Alliance.* 2021;4(11):e202101084. Published 2021 Sep 15. doi: <https://doi.org/10.26508/lsa.202101084>

24. Yue L, Yan M, Chen S, Cao H, Li H, Xie Z. PTP1B negatively regulates STAT1-independent *Pseudomonas aeruginosa* killing by macrophages. *Biochem Biophys Res Commun*. 2020;533(3):296–303. doi: <https://doi.org/10.1016/j.bbrc.2020.09.032>
25. Ye G, Xie Z, Zeng H, et al. Oxidative stress-mediated mitochondrial dysfunction facilitates mesenchymal stem cell senescence in ankylosing spondylitis. *Cell Death Dis*. 2020;11(9):775. Published 2020 Sep 17. doi: <https://doi.org/10.1038/s41419-020-02993-x>
26. Yu C, Zhan X, Liang T, et al. Mechanism of Hip Arthropathy in Ankylosing Spondylitis: Abnormal Myeloperoxidase and Phagosome. *Front Immunol*. 2021;12:572592. Published 2021 Nov 22. doi: <https://doi.org/10.3389/fimmu.2021.572592>
27. García-Salinas R, Ruta S, Chichande JT, Magri S. The Role of HLA-B27 in Argentinian Axial Spondyloarthritis Patients. *J Clin Rheumatol*. 2022;28(2):e619–e622. doi: <https://doi.org/10.1097/RHU.0000000000001763>