Studies in Big Data 114

Siddharth Swarup Rautaray
Manjusha Pandey
Nhu Gia Nguyen *Editors*

Data Science in Societal Applications

Concepts and Implications

Springer

Studies in Big Data

Volume 114

Series Editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland

The series “Studies in Big Data” (SBD) publishes new developments and advances in the various areas of Big Data- quickly and with a high quality. The intent is to cover the theory, research, development, and applications of Big Data, as embedded in the fields of engineering, computer science, physics, economics and life sciences. The books of the series refer to the analysis and understanding of large, complex, and/or distributed data sets generated from recent digital sources coming from sensors or other physical instruments as well as simulations, crowd sourcing, social networks or other internet transactions, such as emails or video click streams and other. The series contains monographs, lecture notes and edited volumes in Big Data spanning the areas of computational intelligence including neural networks, evolutionary computation, soft computing, fuzzy systems, as well as artificial intelligence, data mining, modern statistics and Operations research, as well as self-organizing systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

The books of this series are reviewed in a single blind peer review process.

Indexed by SCOPUS, EI Compendex, SCIMAGO and zbMATH.

All books published in the series are submitted for consideration in Web of Science.

Siddharth Swarup Rautaray · Manjusha Pandey ·
Nhu Gia Nguyen
Editors

Data Science in Societal Applications

Concepts and Implications

Springer

Editors
Siddharth Swarup Rautaray
School of Computer Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, Odisha, India

Manjusha Pandey
School of Computer Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, Odisha, India

Nhu Gia Nguyen
Graduate School
Duy Tan University
Danang, Vietnam

ISSN 2197-6503 ISSN 2197-6511 (electronic)
Studies in Big Data
ISBN 978-981-19-5153-4 ISBN 978-981-19-5154-1 (eBook)
https://doi.org/10.1007/978-981-19-5154-1

Preface

Data science is an amalgamation of various disciplines of research including the method ranging from empirical methods to current trends of data observation along with the application of rigorous skepticism about observations in data, resulting in interprets that involve formulating hypotheses and deducing inductions. These principles of data science have been distinguished in the form of a series of steps applicable to various enterprises. The processes, algorithms, and systems to extract knowledge and insights from a huge amount of structural and unstructured data produced by these enterprises have made data science the need of the hour. The aim is to use the data as raw inputs and generate valuable data products that can be used by enterprises for their business benefits. Data science in societal applications has been established as a new horizon of the scientific field with the silver lining of research evolution in societal applications using renovations in the existing domains of statistics, computing science, and intelligence science. The application of advancements in these legacy research domains and its practical transformation in engineering, public sector enterprises, social science, and lifestyle spheres is the way forward. This adds to the ever-increasing importance of data science, both big data and small data, which when analyzed scientifically bring with them a wealth of opportunities that combine the new trends to extract, transport, pool, refine, store, analyze, and visualize data which are needed to unleash their power. This also simultaneously can be incorporated into tool-integrated development environments and workflows that are ready to be used by people in general.

To have a successful career in data science emphasizing its societal applications for the benefit of the public at large is a real challenge. The desired knowledge base includes complex topics from statistics, computer science, and mathematics. In addition to that, domain-specific knowledge becomes essential for user-friendly application developments. The real-time applications of data science are added challenge that requires to be mitigated. Thus, there is a need for conjugation of data science with various real-life applications of computing and implementations. The proposed book intends to provide a range of current trends in the theoretical and practical applications of data science that are being developed and implemented by researchers, academicians, and industries.

The book covers contributions from more than 13 renowned researchers in the domain who have proven theoretical and foundational research, platforms, methods, applications, and tools in the domain of societal applications of data science. The book will act as a base for future researchers who want to work in the domain of theory and practices in the area of data science. Also, the contents of the book focus on the social, geographical, and temporal dimensions of data science research. This is further supported by real-world examples drawn from a broad range of application areas, including healthcare, bioinformatics, social analytics, and natural language processing as included in the book. The chapters in the book also cover application-oriented perspectives, and advances in the field of data science in such areas as big data collection, searching analysis, and knowledge discovery. This book will be helpful for the students, practitioners, and researchers working in the area of data science and analytics.

Bhubaneswar, India — Siddharth Swarup Rautaray
Bhubaneswar, India — Manjusha Pandey
Danang, Vietnam — Nhu Gia Nguyen

Acknowledgments

The genesis of the book is the 2nd Data Science Conclave 2021 on Data Science in Societal Applications: Concepts and Implications held at Kalinga Institute of Industrial Technology (KIIT), Deemed to be University, Bhubaneswar Odisha, India during October 09–10, 2021 in virtual mode. The invited speakers of the conclave have submitted the full chapters which have undergone a review process for inclusion in the book. Based on the reviewer comments received the authors have made necessary changes in the chapter. We extend our sincere thanks to all these invited speakers for their contributions. We also express sincere thanks to Springer for the support in publishing this book.

We are thankful to the Data Science Group of the School of Computer Engineering, KIIT, Deemed to be University and the whole organizing team of Data Science Conclave 2021 for their confidence in us in shaping up this book as editors. We are thankful to Prof. Sasmita Rani Samanata, Vice Chancellor KIIT DU. who has been a great mover to the Data Science Conclave series since its beginning. We are grateful to the founder of KIIT and KISS Prof. Achuyta Samanta, the patron of the Data Science Conclave series from its inception for the endless support he has provided toward the Data Science Conclave and other events conducted by the Data Science Group of the School of Computer Engineering.

Siddharth Swarup Rautaray
Manjusha Pandey
Nhu Gia Nguyen

Contents

Cloud GIS Model for Coastal Geospatial Big Data Analytics 1
K. K. Barik, Vivek Mishra, J. R. Mohanty, Mrinal K. Debbarma, and R. K. Barik

Appealing AI in Appalling Covid-19 Crisis and the Impending 13
Debanjan Banerjee, Anindita Banerji, and Arunita Banerji

Role of Data Science in Programmatic Advertising 33
Aditya Jain, Kashyap Barua, and Monali Barbate

Social Development Data and Societal Modelling: A Study in Indian Context 47
Rabi N. Subudhi

Deep Learning Trends and Inspired Systems in Natural Language Processing 57
Aseer Ahmad Ansari, Siddharth S. Rautaray, and Manjusha Pandey

AI and IoT Enabled Smart Hospital Management Systems 77
Mahendra Kumar Gourisaria, Rakshit Agrawal, Vinayak Singh, Siddharth Swarup Rautaray, and Manjusha Pandey

Optimization in Agricultural Growth Using AI and Satellite Imagery 107
Bhaskar Goswami and Plaban Nayak

Plenus Solutionem: A Complete Solution to Multi Chronic Diseases Detection and Classification 127
Aindril Kar, Harshit Anand, and Kalipada Tripathy

Kochen Helfer: An AI-Based Cooking Assistant 143
Indrashis Das, Anoushka Mishra, and Shushant Ghosh

Contemporary Digital Consumers: "Real" Versus Virtual Value Co-creation 161
Milena Ratajczak-Mrozek and Marcin Wieczerzycki

Multiclass Prediction of Heart Disease Patients Using Big Data Analytics 177
Sarita Mishra, Manjusha Pandey, Siddharth Swarup Rautaray, and Sabyasachi Chakraborty

About the Editors

Siddharth Swarup Rautaray is an Associate Professor at the School of Computer Engineering, KIIT-Deemed to be University, Bhubaneswar, India. He received his Ph.D. in Information Technology from the Indian Institute of Information Technology, Allahabad, U.P., India. He has over 10 years of teaching and research experience. His research interest areas are big data analytics, image processing, intelligent systems, human-computer interaction, and similar innovative areas. His research contribution includes 06 co-edited proceedings/books, over 110+ research publications in reputed conferences, book chapters, and journals. As an organizing chair, he has organized 05 international conferences and has been part of different core committees of other conferences and workshops. He has delivered invited talks in several workshops and conferences.

Manjusha Pandey is an Associate Professor at the School of Computer Engineering, Kalinga Institute of Industrial Technology, Deemed to be University, Bhubaneswar, India. She has teaching and research experience of over 10 years. She received her doctoral degree from the Indian Institute of Information Technology, Allahabad, U.P., India. Her research areas are big data analytics, computer networks, intelligent systems, machine learning, and similar innovative areas. Her research contribution includes 04 co-edited proceedings/books, over 110+ research publications in reputed conferences, book chapters, and journals.

Nhu Gia Nguyen received Ph.D. degrees in computer science from Hanoi University of Science, Vietnam. Dr. Nguyen serves as Dean of the School of Computer Science at Duy Tan University. His experience includes over 22 years of teaching, and he has more than 40 publications in various conference proceedings and international journals of repute. His research interests include algorithm theory, medical imaging, network optimization, and wireless security. He has been organizing chair of (Springer LNAI) IUKM-2016 held in Vietnam. He is an associate editor in the International Journal of Synthetic Emotions (IJSE) under IGI Global.

Cloud GIS Model for Coastal Geospatial Big Data Analytics

K. K. Barik, Vivek Mishra, J. R. Mohanty, Mrinal K. Debbarma, and R. K. Barik

Abstract Geospatial big data are being shared, analysed, and processed using the Cloud GIS model. When it comes to storing and doing geospatial big data analyses, the Cloud computing paradigm helps to enhance throughput while decreasing latency. In the present research paper, a Cloud GIS Model was presented and constructed for the analysis, sharing, and processing of geospatial big data in India's coastal management, merging cloud computing with geospatial big data efforts. As a result of this model, all of the geospatial data from coastal districts in India would be integrated, making it useful for both decision-makers and planners. While QGIS Plug-in is coupled with Quantum GIS to invoke a Cloud GIS environment, the current model has relied on Quantum GIS for geospatial database development. The model's assumptions have been tested on a subset of the Odisha coast, specifically the Baleswar, Bhadrak, Kendrapara, and Jagatsinghpur coasts.

Keywords Cloud computing · Cloud GIS · Geospatial big data · Odisha coast · Overlay analysis

K. K. Barik
School of Civil Engineering, Centurion University of Technology and Management, Bhubaneswar, India

V. Mishra
Data Engineer, Scanbuy, USA

J. R. Mohanty · R. K. Barik (✉)
School of Computer Applications, KIIT Deemed to Be University, Bhubaneswar, India
e-mail: rabindra.mnnit@gmail.com

M. K. Debbarma
Department of Computer Science and Engineering, National Institute of Technology Agartala, Agartala, India

S. S. Rautaray et al. (eds.), *Data Science in Societal Applications*, Studies in Big Data 114, https://doi.org/10.1007/978-981-19-5154-1_1

1 Introduction

The cloud GIS approach has eased the sharing and exchange of large amounts of geospatial data held by various stakeholders/investors. It has already begun developing an ecosystem that allows a diverse range of users to securely retrieve, store, and transmit spatial and non-spatial data. The expanding technological requirement for data organisation across multiple disciplines and organisations, as well as the requirement for multi-participant and decision-support environments. By eliminating duplication of expenses involved with the development and management of related geospatial dataset, it allows consumers to save money, time, and effort while acquiring new datasets [1–3].

Cloud GIS models can be used for data storage, scaling data and maps to user-defined scales, querying, and analysing the data, and planning/presenting the final map/report to planners/decision-makers [4, 5]. In comparison to desktop GIS computer interfaces, cloud GIS models have achieved considerable appeal and are being used for several reasons. According to Fig. 1, which was developed using the Google Trends analytical environment, during the previous 5 years, cloud GIS has demonstrated a better potential than the desktop GIS computing environment.

The fast development of information technology over the last few decades has resulted in an explosion of data generated by various sensors, bringing us into the era of big data [3, 6, 7]. Geospatial Big data are the emerging field in big data which are containing geographic location information and are generated by sensors such as smartphones and handheld Global Positioning System (GPS) devices have already permeated our daily lives and demonstrate tremendous potential in practical applications such as emergency preparedness and pollution monitoring [6, 8, 9]. Figure 2 illustrates a simple pictorial presentation of geospatial big data that describes the main components.

Fig. 1 Comparative analysis of using cloud GIS over desktop GIS computing environment

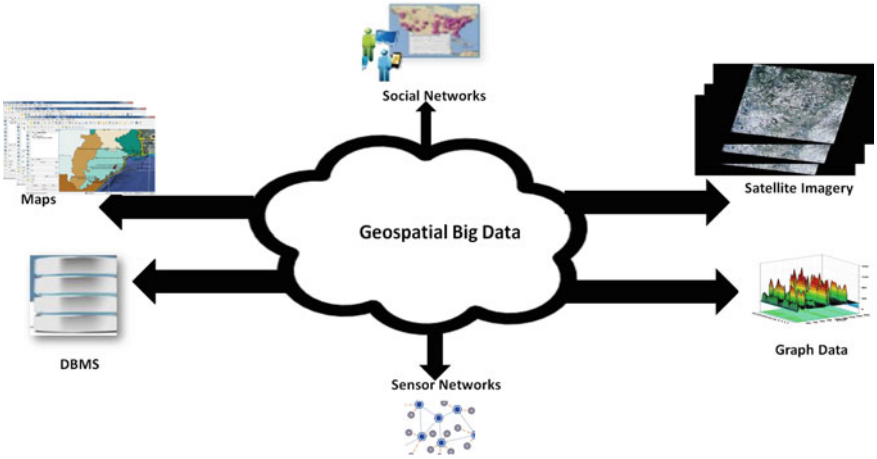


Fig. 2 Components of geospatial Big data

It was possible to process geospatial big data at the edge utilising the Cloud GIS environment in this research report. The subsequent additions to the Cloud GIS Model in a coastal management/administration scenario were made by this paper:

- Cloud GIS model is presented for enhanced throughput and reduced latency for storing and analysing coastal geospatial big data.
- Geospatial data analysis scheme and overlay analysis in thin and mobile client contexts were done.
- A case study of four districts on the Odisha coast, India, was carried out.

The rest of the article is arranged as follows: Sect. 2 provides an overview of pertinent related work. Section 3 highlights the necessity of a Cloud GIS Model for the Coastal Information Infrastructure Network. The purpose of the current investigation is presented in Sect. 4. Section 5 outlines the suggested Cloud GIS Model, whereas Sect. 6 expounds on the results and discusses the produced model with overlay analysis. Finally, Sect. 7 provides the current study's concluding remarks and future recommendations.

2 Related Works

The cloud computing environment has given a huge computational resource for various geographical analysis. The cloud computing model allows users to move from desktop to web server. Cloud computing and additional web processing frameworks have enabled asset sharing over the web. The Cloud GIS Model connects enterprises with technology, tools, and people to produce, handle, and use geographic statistics and data [5, 10, 11].

Similarly, Cloud assisted GIS Model has a single-instance, multitenant approach, allowing several clients without interfering with one another. This integrated hosted service technique aids in the installation of patches and application improvements for the benefit of the user. Another feature of Geospatial Cloud is its acceptance of SOA and numerous web services, a completely recognised architectural paradigm in service engineering. Numerous Cloud systems expose the data and functionalities of apps through web services [9, 12]. Figure 3 depicts the Cloud GIS Model's system architecture with three critical components.

For visualising spatial data, the client tier layer includes thin, thick, and mobile clients' environments. Mobile customers are people who use mobile devices. Thin clients are users who work on simple web browsers. Users don't need any additional software to use thin clients [4]. Thick clients are users who process or visualise spatial data in independent systems that require extra software to operate fully [13]. The Application Tier includes server-side geospatial services. It acts as an intermediary between clients and providers. In order to provide various services such as Web Map Services (WMS), Web Coverage Services (WCS), Web Features Services (WFS), Catalogue Services for Web (CS-W), and Web Processing Services (WPS),

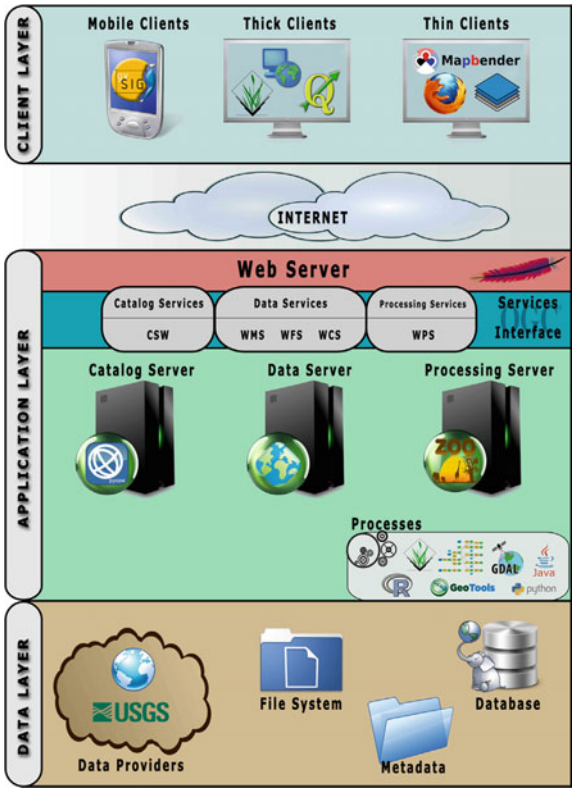


Fig. 3 Cloud GIS model with client, application and data layer for geospatial data sharing, analysis and storage [12]

a dedicated application servers have been maintained at the top of the application tier [12, 14].

OGC WPS standards are used in several open-source projects [2, 5, 15]. Quantum GIS is utilised for numerous web services in the cloud [16]. Flexible, ascendable, and sustainable environment provided by open-source lightweight Docker technology in cloud computing containers with Grass GIS, GDAL, NumPy, and IPython. Several cloud-based mobile design architectures Apps for data collection and geographic data visualisation have been created for Android. For geographical data visualisation and collection, mobile/smart device computing with cloud-based systems has shown fantastic promise. Grid, vector and maps and remote sensing images are required for data processing and surveying in forest inventory management. To control pollutants, air quality and greenhouse gases, the California and US EPAs have created an open-source smartphone app [14, 15, 17].

Geographic web services, geospatial database generation, and geospatial modelling services are necessary for the construction of a seamless Cloud GIS Model. Thus, it may leverage open-source GIS tools to implement these services. Thus, by integrating cloud computing with GIS technology, the urgent requirement is to build an easily perceivable and usable coastal information network for the general public in India. This should also meet a long-awaited requirement for those who want to continue their research on coastal administration/management. The network's transparency should also exert pressure on decision-makers to guarantee sound coastal management. This has demanded the creation, development, and deployment of a Cloud GIS Model for a coastal information network [14, 15].

3 Need for Cloud GIS Model for Coastal Information Infrastructure Network

The convergence of cellular phone, web, and geospatial technology has increased the capability for a wide range of functionality in conditions of geospatial data exchange over digital cyber space. It may be able to provide a dynamic and real-time manner of representing valuable information/data using maps. As a result, there is an immediate need to build a well-organised Cloud GIS Model that acts as a robost intuitive geoportal for all stakeholders to utilise, share, and access geospatial data for socio-economical and environmental purposes. Geospatial web services are the main critical technology for the development and deployment of the Cloud GIS Model. Cloud computing technology is employed in the design and execution of the Cloud GIS model, which is used to share information about the coastal information specifics of four districts in Odisha, India. It enables data analysts and inexperienced users to swiftly investigate a problem and obtain the information they require. As a result, the next section discusses the specifics of relevant work done with the Cloud GIS Model.

4 Objective of Present Study

After reviewing numerous relevant research, the purpose of this research is to build as well as implement a prototype constructed on the Cloud GIS Model for India's Coastal Information Infrastructure Network. Additionally, it presented the cloud GIS Model's system design, focusing on the web browser and mobile environment. Additionally, it presented a sound incremental procedure for developing the integrated geospatial and non-spatial database for the Coastal Information Infrastructure Network using Quantum GIS Ver 3.14. The current study employed a case study approach, focusing on four districts in Odisha, India. Additionally, it has compiled district-specific geomorphological data. The research area is a 283-km-long coastline length on India's east coast, comprising chosen regions of Odisha, namely, Baleswar, Bhadrak, Kendrapara, and Jagatsinghpur, adjacent to the Bay of Bengal, as illustrated in Fig. 4.

The western end of the research area is well-drained by major rivers like the Subarnarekha, Budhabalanga, Baitarani and Mahanadi, which provide the study region with the majority of its sediment. The mangrove forest of Bhitarakanika is a popular eco-tourist destination. The study area is bounded by the ports of Paradeep and Dhamara. It is located between latitudes 17 0 49′ N and 220 34′ N and longitudes 810 27′ E and 870 29′ E. It is despicable that these coastal districts of Odisha regularly

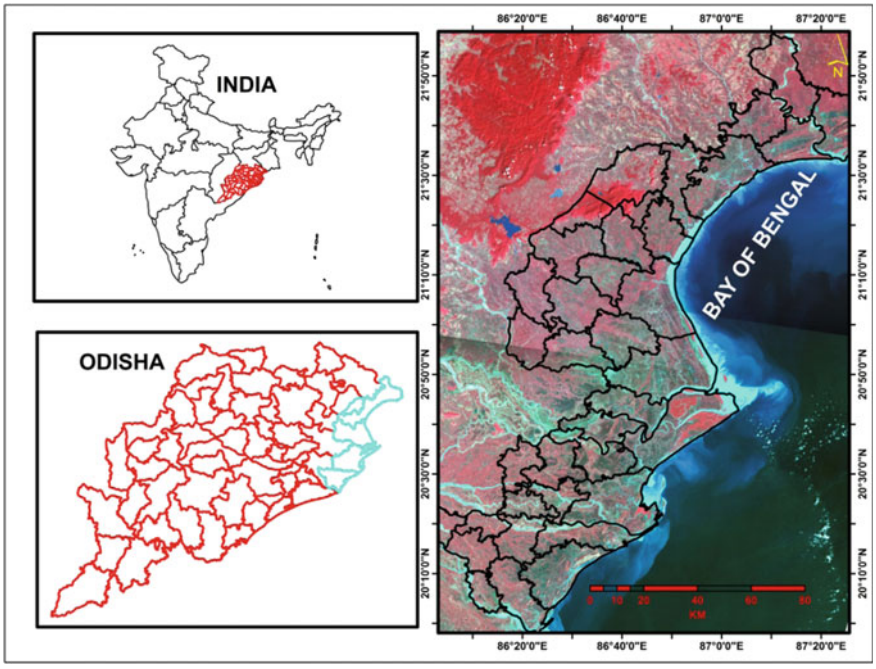


Fig. 4 Location map depicts the study area

lose lives and sustain serious damage due to a variety of coastal hazards, including severe riverine flooding, tropical cyclones, and coastal storm surges originating in the Bay of Bengal. Thus, the next part explains the Cloud GIS Model's suggested system design.

5 Proposed Architecture of Cloud GIS Model

Three critical layers comprise the architecture of the Cloud GIS Model. The layer's initial phase is the data tier layer. Geospatial databases have been generated in this layer using Map Window GIS and Quantum GIS. Numerous data sources, such as Google WMS and Open Street Map, were combined with the generated database for additional study. Catalogue services have been kept at the application layer for the purpose of searching for needed services. For the purpose of invoking WFS and WMS services, a dedicated data server is maintained via the QGIS Cloud Server Provider. The QGIS Cloud Service Provider is ultimately accountable for application tier layer maintenance. Ultimately, the client tier layer is used by the three groups of consumers. The Cloud GIS Model's suggested system architecture is depicted in Fig. 5.

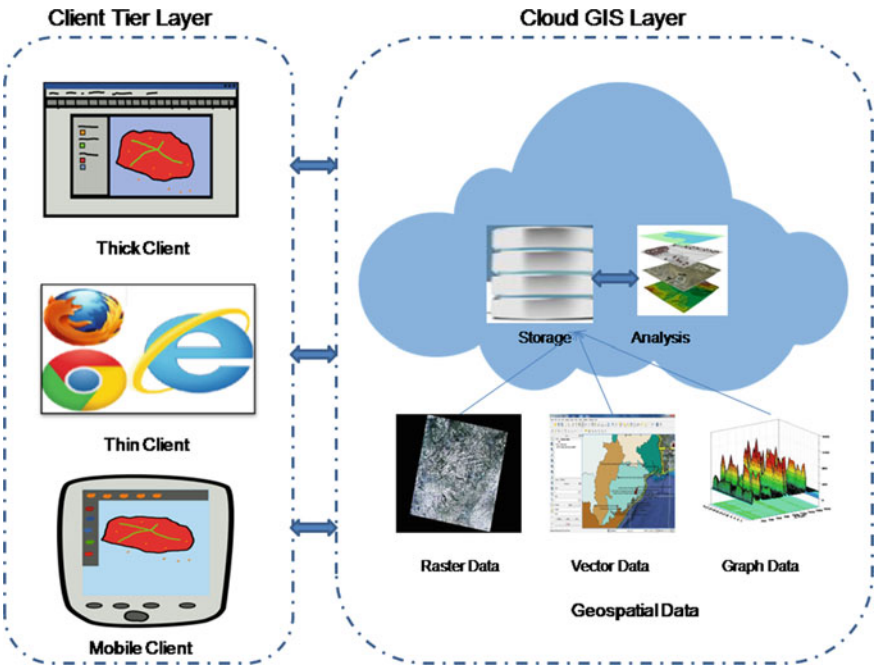


Fig. 5 Proposed system architecture of cloud GIS model

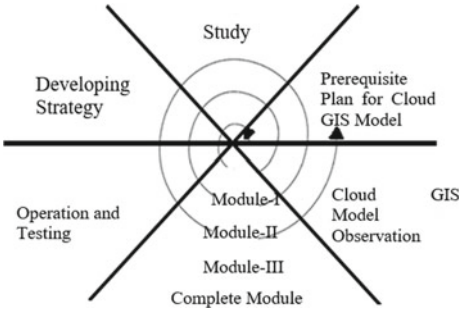


Fig. 6 Approaches to prototype development based on the spiral model

5.1 Methodology Implemented

The construction of geospatial databases has mostly focused on a practical method of discovering and spreading the notion. To use the Cloud GIS Model, a geographic database must be developed. Figure 6 shows how to create a geographic database using a spiral model.

This will allow research of the cloud computing interfaces. The suggested Cloud GIS Model has three main functional parts. For the Coastal Information Infrastructure Network in India application cases, Module I comprises the geographic database created in GIS context. Assembling OGC WMS/WFS compatible geospatial web services is the subject of Module II. It also explains how these services work with QGIS Cloud. How to integrate the Model with different sorts of users is covered in Module III. To fully develop the Cloud GIS Model, Open-Source GIS should be used optimally. Each stage also aims to reveal specifics about the underlying infrastructure-based web services.

5.2 Module Integrations

The creation of geospatial databases is a regular or common occurrence. Through the evaluation and testing of a complete module, each iteration phase enhances the study and approach steps. Quantum GIS is used to construct the database for the Coastal Information Infrastructure Network in India using the Survey of India toposheet. Quantum GIS ver. 3.14 is employed for the building of an integrated geospatial database for the Coastal Information Infrastructure Network of four districts in Odisha, India, as shown in Fig. 7. WGS-84 with the EPSG:4326 coordinate reference system is utilised as the universal coordinate system in the generated geospatial database.

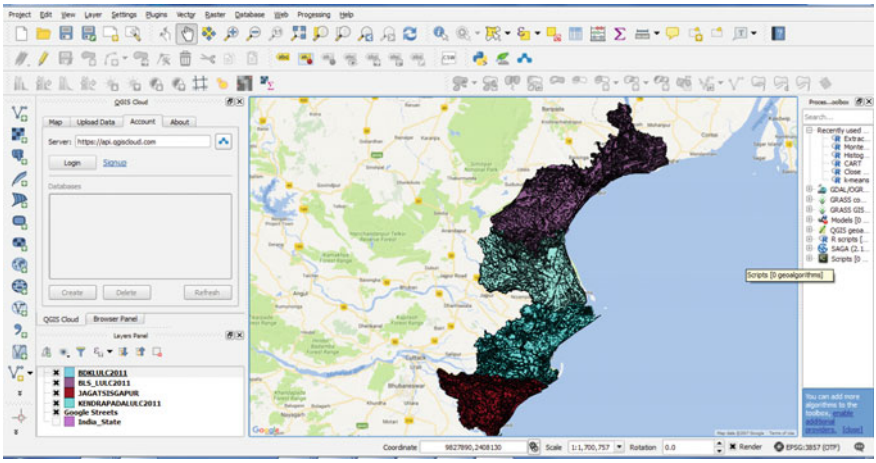


Fig. 7 Integration of coastal geospatial database

6 Results and Discussions

In this section, the overlay analysis is implemented with superimpose of multiple geospatial data on a common platform for better analysis by geospatial data analysts using raster and vector geospatial data. It is observed that the overlay analysis of raster and vector dataset from a certain area is conducted. Initially, the created geographical databases are accessed in Quantum GIS and certain joint operations were conducted. The intended overlay procedure was completed with a web browser, which is known as a thin client operation. A plugin called QGISCloud has been installed in Quantum GIS. The aforementioned plugin can save multiple vector and raster data in a dedicated cloud database for subsequent overlay analysis. Figures 8 and 9 depict the overlay operation in the Cloud GIS Model on a thin client and a mobile client environment, respectively.

7 Conclusions

The Cloud GIS Model that leveraged a gateway for coastal geospatial big data analytics was designed and validated in the current research article. The overlay analysis has been completed with the incorporation of Google WMS in the current working model. The proposed model has a lot of promise for integrating all of India's coastal states into a single integrated platform and sharing coastal information infrastructure. The geospatial database of four districts in Odisha was used in this study. In a geospatial cloud environment, this model also introduces edge intelligence. The proposed model's middle layer will be enhanced in the future with the addition of more intelligent processing. It intends to develop this model in an edge and fog

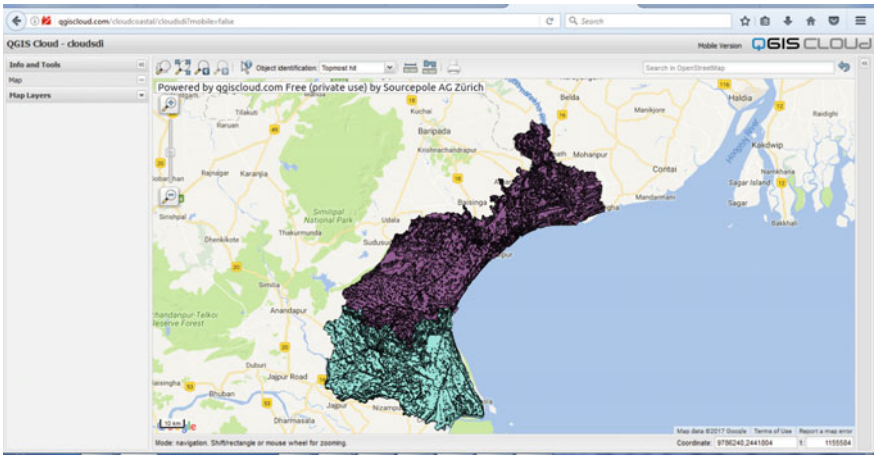
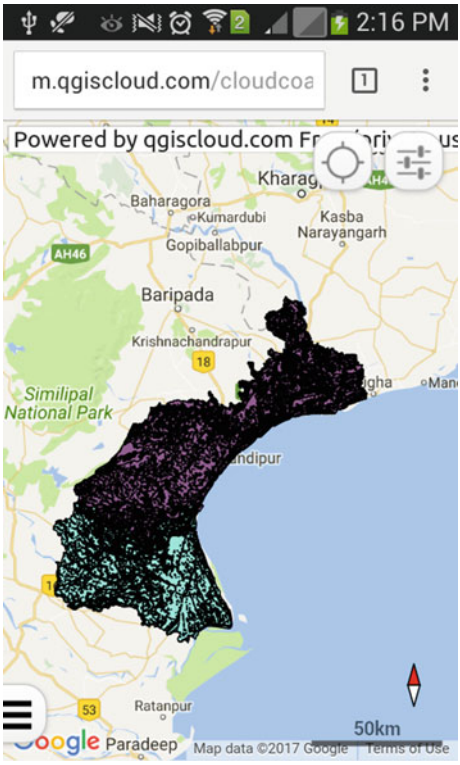


Fig. 8 Integration of Odisha's coastal information infrastructure network in a thin client environment [18]

Fig. 9 Integration of the coastal information infrastructure network of Odisha, India, using mobile client environment [19]



computing environment, and it is likely to add more geospatial web services features as its geospatial database expands across India's coastline states.

References

1. Georis-Creuseveau, J., Claramunt, C., Gourmelon, F.: A modelling framework for the study of spatial data infrastructures applied to coastal management and planning. Int. J. Geogr. Inf. Sci. **31**(1), 122–138 (2016)
2. Pettit, C., Stimson, B., Barton, J., Goldie, X., Greenwood, P., Lovelace, R., Eagleson, S.: Open access, open source and cloud computing: a glimpse into the future of GIS. In: Handbook of Planning Support Science. Edward Elgar Publishing (2020)
3. Cai, P., Jiang, Q.: GIS spatial information sharing of smart city based on cloud computing. Clust. Comput. **22**(6), 14435–14443 (2019)
4. Barik, R.K.: CloudGanga: cloud computing based SDI model for Ganga River Basin management in India. Int. J. Agri. Environ. Inf. Syst. (IJAEIS) **8**(4), 54–71 (2017)
5. Tripathi, A.K., Agrawal, S., Gupta, R.D.: Cloud enabled SDI architecture: a review. Earth Sci. Inf. **13**(2), 211–231 (2020)
6. Deng, X., Liu, P., Liu, X., Wang, R., Zhang, Y., He, J., Yao, Y.: Geospatial big data: New paradigm of remote sensing applications. IEEE J. Select. Topics Appl. Earth Observ. Remote Sens. **12**(10), 3841–3851 (2019)
7. Senaratne, H., Mobasheri, A., Ali, A.L., Capineri, C., Haklay, M.: A review of volunteered geographic information quality assessment methods. Int. J. Geogr. Inf. Sci. **31**(1), 139–167 (2017)
8. Wang, S., Zhong, Y., Wang, E.: An integrated GIS platform architecture for spatiotemporal big data. Futur. Gener. Comput. Syst. **94**, 160–172 (2019)
9. Yang, C., Huang, Q., Li, Z., Liu, K., Hu, F.: Big Data and cloud computing: innovation opportunities and challenges. Int. J. Digital Earth **10**(1), 13–53 (2017)
10. Slocum, Z., Tang, W.: Integration of Web GIS with high-performance computing: a container-based cloud computing approach. In: High Performance Computing for Geospatial Applications, pp. 135–157. Springer, Cham (2020)
11. Zhang, J., Xu, L., Zhang, Y., Liu, G., Zhao, L., Wang, Y.: An on-demand scalable model for geographic information system (GIS) data processing in a cloud GIS. ISPRS Int. J. Geo-Inf. **8**(9), 392 (2019)
12. Evangelidis, K., Konstantinos, N., Stathis, M.: Constantine, Papatheodorou 'Geospatial services in the cloud. Comput. Geosci. **63**, 116–122 (2014)
13. Yue, P., Zhou, H., Gong, J., Hu, L.: Geoprocessing in cloud computing platforms–a comparative analysis. Int. J. Digital Earth **6**(4), 404–425 (2013)
14. Mantas, V.M., Liu, Z., Pereira, A.J.S.C.: A web service and android application for the distribution of rainfall estimates and Earth observation data. Comput. Geosci. **77**, 66–76 (2015)
15. Gkatzoflias, D., Mellios, G., Samaras, Z.: Development of a web GIS application for emissions inventory spatial allocation based on open source software tools. Comput. Geosci. **52**, 21–33 (2013)
16. Leidig, M., Teeuw, R.: Free software: a review, in the context of disaster management. Int. J. Appl. Earth Obs. Geoinf. **42**, 49–56 (2015)
17. Dong, C., Liu, F., Wang, H., Chen, F.: Application research of mobile GIS in forestry informatization, In: Computer Science and Education (ICCSE), 2010 5th IEEE International Conference on, (2010): 351–355
18. Internet-1: http://qgiscloud.com/cloudcoastal/cloudsdi. Accessed 10 May 2021
19. Internet-2: http://m.qgiscloud.com/cloudcoastal/cloudsdi. Accessed 10 May 2021

Appealing AI in Appalling Covid-19 Crisis and the Impending

Debanjan Banerjee, Anindita Banerji, and Arunita Banerji

Abstract The outbreak of the deadly Covid-19 virus has snatched smiles from everyone's face and now the entire world has been affected directly or indirectly by the effects of the virus, this virus keeps on mutating due to which there is no proper medicine or a final vaccine that assures it will curb the spread of the virus, major countries all over the world has lost more people than in a war and is still losing its people even after getting fully vaccinated. The horror is so much imbibed in each human it seems unrealistic to even think that the world will be normal ever again. This outbreak of the unknown virus is certainly a black-swan event that has annihilated people economically, emotionally, and socially and has made each individual realize the importance of one's health and how to be a responsible person by taking care of whatever finances one has, as in unprecedented times savings are the only resort left with a person. It is a testing time and everyone is at war, we all are soldiers in this pandemic and our health care workers, administration, and government are trying their best to stop the spread of the disease as it has killed more than four lakh people in India only and in the world tally is more than forty lakhs with numbers increasing. In this appalling situation when everything has been shifted to online mode solutions must be looked at in more technologically driven methods, in today's world due to rapid advancement in the IT and computer science sector there are ways to track the next rising hotspot of the virus and how it can be contained by taking swift actions if predicted within a particular time frame. Data collection, data analysis, and studying trends can help in assessing the upcoming threats, and in this manner, new job opportunities can also be created as it will involve people being prepared with limited medical knowledge to cure the people affected with the virus. In these times government and administration must adopt technologically backed solutions

D. Banerjee (✉)
Kalinga Institute of Industrial Technology (Deemed to be University), Bhubaneswar, India
e-mail: debanjan.banerjee98@gmail.com

A. Banerji
Tata Institute of Social Sciences, Guwahati, India

A. Banerji
FORE School of Management, New Delhi, India

S. S. Rautaray et al. (eds.), *Data Science in Societal Applications*, Studies in Big Data 114, https://doi.org/10.1007/978-981-19-5154-1_2

which will help the system to make accurate decisions based on real-time data-driven modeling capable of identifying the relevant information.

1 Introduction

This chapter will explore how AI can help in tackling pandemic-like situations with prior analysis by collecting relevant data, identifying the underlying patterns of influence and giving the most appropriate results after analyzing the suitable solutions with reference to the present dynamics of events along with the historical trend relevance. AI can be used very diligently to identify the virus hotspot areas, what are the trends in recent times, and how they can be tackled including various other factors like how people are following the protocols and how people are getting vaccinated, whether vaccination is helping in bringing down the fatality rate or chances of spread of the virus is decreasing; this all can be estimated with a set of models which work on relevant data which uses past events to give predictions using probability functions induced with complex mathematics.

It is not a hidden fact that the second wave of Covid-19 ripped India apart as the administration was not ready nor there were proper protocols followed plus a crisis in health care units and hospitals was seen, this was due to unpreparedness and too much relaxation of the protocols. In such a situation AI comes into play as it can analyze tons of data and predict what can arise in the future. The predictions based on recent dynamics can give results that are not too far away from the truth, with prior information and tracking of people the administration can take required steps to control the situation and the health care system can be alerted so that they can be prepared with enough resources. In this chapter, various tools and models are discussed which can be used for future predictions using machine learning, statistical models, and complex mathematical programs. The end goal of the proposed framework is to build an AI-powered Autonomous Assistance Algorithm which will proactively not only help the governing authorities in tackling pandemics like Covid-19 but will be playing a key role in assisting humanity to fight back to normal using its intelligence.

2 Covid-19: The Key Observations and Learnings

In this unfortunate and unprecedented situation, the entire world has faced immense disruption and humongous loss of lives. Not a single country was prepared for such a situation where everything seems out of control and the entire administration and health care system fell apart, ripping the countries to their fullest extent. The virus is attacking in waves and is mutating thus, no one is sure about the third wave, many experts say that it is inevitable but its spread can be curbed if required measures are taken on time and the system is prepared for it.

In the past two waves the key observations and learnings are as follows:

- Administration Challenges:
  The entire situation was handled based on how other countries were handling it, India's huge population and the existing nature of work life was not ready for such a shock, the government ministries failed to detect the source of the virus and were not able to identify the viable hotspots before the cases in those particular areas reached its peak. The unpreparedness of the administration can be seen in every state, the second wave was more catastrophic and it could have been like this if the administration had taken enough measures to create more healthcare facilities and had banned all the gatherings for and year after understanding the sources and hotspots of the virus. Due to a lack of efficient management the doctors were also not provided with the advanced requirements that were advised, they had to provide their services with limited resources. There were no mandatory checks at the airport and the incoming passengers from countries were not entitled to mandatory quarantine for fourteen days as well, this all exposed the residents of that state exposed to the virus in an unimaginable manner. Once the situation started worsening then only lockdown and bans were issued, the virus cannot be stopped from spreading by taking spontaneous actions it can be stopped by imposing precautionary measures in its toughest forms.
- Unemployment:
  One of the adversely affected is the middle-class youth, who had to undergo immense pressure and have the responsibility of their families. In this situation, many young adults lost their family and friends hence making them the sole breadwinner and this compelled young students to quit education and start searching some jobs to support their families but not many were successful in their endeavors, the world went through mass-layoffs and may employees who were in heavy machinery, etc., lost their jobs and had a family to look after, they were compelled to do petty jobs. Due to a sudden halt in all the day-to-day work, many migrant laborers lost their jobs, had no source of income, faced severe food shortages, and were uncertain about the future. Government had made provisions for them, but those provisions did not reach them in a systematic manner. These workers then wanted to go back to their villages as in several rural areas the spread of the virus was less and farming along with other livelihood work were still going on. The crisis reached its peak when migrant workers were trying to go back to their homes which created a lot of crowding in bus stations and railway stations.
- Impact on Academia:
  In the wake of the Covid-19 pandemic, the entire education system has shifted to the online mode which has exposed the students, teachers, and educators to a pool of challenges. Many students do not have the proper resources or live far off to catch good signal strength, it has become immensely tiring for the students to attend lectures online and study at home all by themselves with some help from their parents. There is no personal attention toward the weaker students and in rural parts, the online mode of education still has a long way to go to be reachable to the students, during this pandemic 11% of the rural households have

bought smartphones but cannot access many things. The students who have all the facilities also cannot perform well in practical subjects and the entire evaluation system had to change in a course of the year. Students pursuing bachelors and masters have been hit the worst, as in colleges there are students from various parts of the country. Students cannot always be available and face immense connectivity problem, from online class presentations to strenuous online exams it has become difficult for the students to be able to complete their academic commitments. Students who have changed their streams are clueless and do not know how to cope up with the syllabus and whom to ask, there have been no or minimal senior-junior interaction which possess even more challenge for the students. As masters or bachelors are the base of the professional world, in this if the students are not getting the proper exposure and experience the recruitment rate will fall eventually, currently only those organizations are hiring where work can be done in online mode for other practical exposure courses there are very less opportunities. For example, students pursuing masters or bachelors in social work are not getting the opportunity to work in the grass root levels and have lost many job opportunities as well, organizations working for these causes are now funding the pharmaceutical companies and vaccine drives leaving other aspects void for the moment. In these trying times, even the teachers are exhausted as they are not able to teach the way they used to and are facing challenges coping up with the digitized way of teaching, many school and college teachers are not aware of the technological tools in many online teaching platforms and they also face immense power cuts and connectivity issues, especially when they teach from their hometown(s). It has become challenging and tiring to evaluate the students and grade them in online performance as it is not easy to understand how much the student has actually been able to follow and whether the work is authentic or not. Female educators have to look after their families and teach their own children as well, responsibilities of family and school/college come together for these educators which makes it tougher for them to deliver their best as well. Thus, the online mode of education can be a temporary solution but continuing like this will weaken the base of the students and they cannot ever learn practical subjects or get field exposure that helps them in securing suitable jobs.

- Health Implications:
  Prolonged screen time and no or no or minimal exercise have exposed students to various health problems and most of the students are getting higher power glasses as well, entire work-from-home concept has caused devastating health effects on everyone. The youngsters are getting back pain and spinal cord pain which is not appreciated at such an age. Due to no physical activity school-going students and adolescents are getting hyperactive which is causing them to behave in certain abnormal ways, female homemakers have faced a lot of pressure once everything became online as they had to take care of everyone in the home from elderly parents to tender toddlers without any external help, the entire day's work caused numerous health issues as well like fatigue, body pain, and muscle pain, in such situation it is not possible for only one person to carry out all the household chores. People sitting in front of the screen all day long face problems like dizziness,

headache, vision problem, mild throbbing in a particular part of their head, and causing multiple pain areas. It is taking a toll on the people involved in office work which has increased due to the pandemic and the pressure to survive in this cut throat competition too is hounding the employees for performing well. From prolonged timings to early deadlines it has become very difficult to be able to keep track of health and other ailing issues; due to all these people have been neglecting their persisting health issues and have not been able to go to doctors for regular check-ups as well.

- Depleting Mental Health:
  Mental health now in the aspect of life is depleting and from toddlers to adults no one has been spared from the havoc. People all over the world are going through tremendous mental pressure and in this situation it has become impossible to maintain mental peace and carry on with everyday tasks. Situation is so bad that people have gone to the cremation center finding no vacant slot and stood alone with their loved ones' deceased bodies. Humanity has witnessed so many deaths and Covid-19 cases which has made everyone helpless and has forced them to live in negativity. From loss of lives to loss of jobs, from ill-maintained medical infrastructure to a mockery of education, the pandemic has made us through everything. It has become practically impossible to think that life will be normal ever again. Students cannot go out and study they have to rely on computers and laptops only, working professionals cannot assess the real ground level work which is affecting the people under them apart from IT sector professionals. It has all become just a virtual world where human touch is now scared of and fear of getting exposed to the virus is always there in the hearts and minds of everyone. Women homemakers are also not spared from this nightmare, in this situation they have to take care of their families and in most Indian families aged and children live which becomes a challenge for the female as she has to do all the household chores, house-help being unavailable and taking care of aged parents and school-going children. First wave made us all worry about the future but the second wave ripped us all apart. Few states in India were in such a situation where one in every three tests was positive and all the members of the family were affected by the virus. Seeing people dying in front of one's eye is not a thing to process easily, mental toll today is more than a war, we all are in the battlefield even being at home. People in rural India are now out of the communication circle which leaves them at the edge and they don't have the required medium of communication as well, it has made them vulnerable and without any income, as there still is inter-state travel restrictions they cannot come to big cities to resume their work and earn their daily wage, they cannot undergo state-mandated quarantine as the cost has to bore by the people traveling only. Pandemic sure is a black-swan event that has put a halt on everyday life and only the most privileged are unaffected and indifferent. Students, adolescents, young professionals, young mothers, and elderly people from every walk of life have been affected mentally and cannot live the way they had wished. Every aspect of life has been stopped in some or other way students and young professionals are committing suicides more than ever which is an alarming matter to look into, lack

of availability of resources and equipment have forced many workers to search for meager jobs which barely support their families. In this constant state of confusion and uneasiness life has become extremely challenging for small business owners, street hawkers, restaurant owners, hospitality sector, tourism, and the list never ends. Anxiety, helplessness, trauma, depression, unhappiness, and confusion has become synonymous with one's survival nowadays, people who have recovered from the disease face issues like extreme heart palpitation, indigestion, headache, and other issues to which doctors also have no proper solution, such a state of mind leads to anxiety and tension which further depletes one's mental health. Solution to a disease is a drug/medicine, vaccination boost the immune system in a body but does not guarantee relief from the disease, people are so much disturbed and disheartened that they fear to discuss their issues with anyone and feel they can handle everything on their own which technically is not possible. In these trying times, people should stand for each other in any way possible so that no one loses hope and keep on trying their best until the black phase is finally over.

3 Proposed Solution

The objective of this solution framework is to build an Artificial Intelligence-enabled robust digital framework that is capable of collating different sources of information for forecasting the next wave at a more granular level. This will help the nation to effectively combat the challenges and unprecedented threats posed by this invisible pandemic war. This pandemic has not only impacted the population through the virus incursion but also due to economic and mental collapse, the developing countries are suffering from unemployment and hunger crises. Our solution framework will also help in identifying the best suitable employment opportunities for the impacted individuals.

Today's state-of-the-art forecasting tools use machine learning, a type of AI that relies on historical data to make predictions for the future. But unlike recurring epidemics, which provide useful information in their wake—pandemics offer little historical data to learn from. By definition, a pandemic is the worldwide spread of a new disease. That means that, in the beginning, there is no availability of historical data which can be leveraged to build and train the model. To tackle these challenges, we propose an Autonomous and incremental artificial intelligence framework that will be capable of adapting to the new behaviors using multivariate point process modeling. Our underlying hypothesis behind this experiment is, that when the events occurring over a certain period of time are stochastically inhibited, the sequence of events will be playing a crucial role in deciding the likelihood of the future event. The point process modeling is highly flexible in accommodating demographic features along with mobility trends, which are connected using statistical models. This will help us in forecasting the trend at a more granular level. This will help the authorities to ***take*** proactive pivotal measures in advance to combat the fight against Covid-19, by scaling up the medical care facilities and optimizing the logistics effort. While

uncertainties and changes are an inescapable part of this pandemic, our proposed framework will be playing the role of a helping aid in this catastrophic pandemic management from here to eternity.

Every year, more than 120 million workers across India migrate from rural areas to larger cities for employment opportunities. During the lockdown, these migrant workers who wanted to return home often struggled for food, shelter, transport, and employment opportunities. Although humanitarian aids and supporting offers from government agencies, non-profitable enterprises were offered from every corner of the country to meet their most-pressing needs, connecting the help providers with the people in need at scale seems to be the biggest challenge.

Our proposed framework is capable of enabling the unemployed workers community to locate the resources closest to them using machine learning algorithms streaming real-time information at scale. The algorithm attempts to find the optimal similarity matrix between the workers and the openings available through semi-supervised learning using KNN-based Label propagation and Label spreading approaches and adaptive PU learning. The algorithm further adds a layer of regularization to be more robust toward the noise.

4 Solution Architecture

The proposed architecture consists of five layers stacked over one another. Layer one will be responsible for accumulating all sources of information arriving from the governing authorities. The next layer contains AI-powered point process algorithm. The algorithm acts as the backbone of the framework. It helps us analyze every single information feed arriving into the system with utmost granularity. Layer three acts as the response layer, forecasting the next hit and assisting in proactive measures of precaution and safety. Layer four is an extensive wing of the framework specially designed for tackling sociological challenges faced by humanity during this pandemic. This pandemic has taken a big toll on daily wage earners, the unprivileged, and the homeless. The machine will be acting as a helping hand to those in need using optimized nearest neighbors propagation and utility spreading approaches at scale.

The concluding layer holds the responsibility of analyzing the driver along with monitoring and vigilance, which will be backpropagated to the layer two and layer four algorithmic framework for incremental learning and proactive refinements improving the data-centric decision driving capabilities.

The framework architecture of the proposed solution can be found in Fig. 1. Diverse sources of data will be considered for accumulating feeds such as feeds from the governing authorities regarding different demographical information, domestic and international border activities, social media extracts, hospital occupancy data, travel itinerary, extracts from GPS, Smart devices, WIFI, etc. These information feeds will be further processed with sparsity prediction/removal, outlier management, and different data wrangling processes using state-of-the-art exploratory data analysis and cleansing algorithms. Next, they will be stored in data centers in-house/cloud

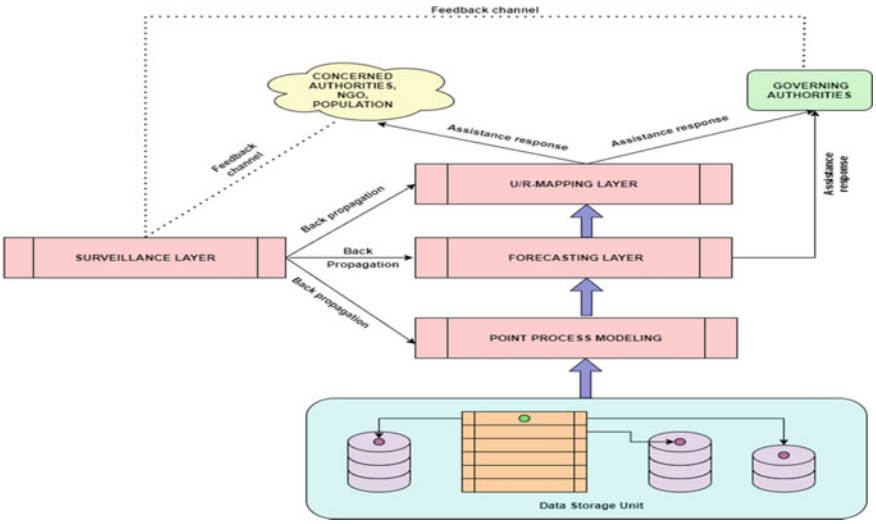


Fig. 1 Proposed Solution Architecture

storage. These storage units are capable of storing heterogeneous data, which can be employed to store any type of data objects.

4.1 Point Process Modeling: Hawkes Process

Hawkes process was first introduced by Alan Hawkes in 1971. It is a statistical modeling pipeline, build on self-exciting methods where the probability of an event's occurrence is dependent on the likelihood of another event's occurrence. This basically shows how on one particular event other events are dependent and follows certain patterns to give the most appropriate result, it is a process which is widely used to assess the occurrence of crimes due to various factors, health of people and susceptibility to diseases, market analysis and in various other fields. It helps to predict future actions and how to control certain disasters by taking appropriate measures beforehand.

Hawkes process is a statistical framework for building self-exciting modules, used for building the pipeline to predict the trend and then make users or authorities alert for the future course of action, it follows certain data and gives the most appropriate result, there are few Hawkes Models which are used to solve the complex situations and it helps in accurate prediction of the scenario when provided with required data.

Understanding Various Hawkes Processes in Prediction of Covid Hotspot Generation:

- Hawkes Model
  The intensity function of the Hawkes pipeline is defined as:-

$$\lambda(t) = \mu + \sum_{j:t_j<t} \psi\left(t - t_j\right) \tag{1}$$

  In the above equation, μ is the base intensity and $\psi(\cdot)$ is a pre-specified decaying function, i.e., exponential function and power-law function. Naturally, Eq. 1 means that each of the past events of contraction with Covid-19 virus has a positive contribution to the occurrence of the current event of generation of a new hotspot area, and this influence decreases with the passage of time.
- Intensity Function:
  The intensity function [1] is assumed that all the recurring events associated with a number of cases and other underlying events affect the probability of the formation of a new hotspot. This helps in leveraging the information of the historical and co-occurring events like incoming travelers, migrant labors, people who contracted with the virus that have been updated/reported. The intensity function is considered to be representative of the baseline intensity and the densities of different the historical or co-occurring events of contraction with the virus which might lead to another hotspot generation.

$$\lambda_{y,m}(t) = \mu_y + \sum_{t_\ell<t} \mathbb{I}(m_\ell = m)\alpha_{y_\ell,y}\kappa(t - t_\ell) \tag{2}$$

  Here the first term represents the constant base intensity of generating label y. The second term represents the influence of the events and the cases reported that happen prior to the time of interest. The influence from each event's occurrence decays over time and is modeled using an exponential decay term $\kappa(t - t_\ell) = \omega \exp(-\omega(t - t_\ell))$. The matrix α of dimensionality $|Y| \times |Y|$ represents the encoding of the level of dependency between pairs of areas belonging to the events related to getting contracted with the Covid-19 virus, e.g., questioning an area's trend (increasing/decreasing) may influence the chances of rejecting situation in the future generation of a hotspot area differently from how it would influence a new area where more people are traveling.
- Likelihood Function:
  The different set of parameters present in the intensity function gets optimistically decided by maximizing the likelihood of the events of contraction with the virus. The complete likelihood function is given by

$$L(t, y, m, W) = \prod_{n=1}^{N} p(\boldsymbol{W}_n \mid y_n) \times \left[\prod_{n=1}^{N} \lambda_{y_n,m_n}(t_n)\right] \times p(E_T) \tag{3}$$

Here the first term provides the likelihood of generating a spike in the number of cases given the area and is modeled as a multinomial distribution [2] conditioned on the area.

$$p(\boldsymbol{W}_n \mid y_n) = \prod_{v=1}^{V} \beta_{y_n v}^{W_{nv}} \tag{4}$$

Here V is the population sample size and β is the matrix of size |Y | × V specifying the point process model for each label. The second term provides the probability dynamics in the number of cases at times [t_1, …., t_n] and the third term provides likelihood in a way that the probability density curve flattens at [0, T] timestamp, with the exceptions incurred at times [t_1, …., t_n]. The parameters of the model get optimized by the log-likelihood maximization.

$$\begin{aligned} l(t, y, m, W) = &- \sum_{y=1}^{|Y|} \sum_{m=1}^{|D|} \int_0^T \lambda_{y,m}(s) ds + \\ &\sum_{n=1}^{N} log \lambda_{y_n, m_n}(t_n) + \sum_{n=1}^{N} \sum_{v=1}^{V} W_{nv} log \beta_{y_n v} \end{aligned} \tag{5}$$

The term of integration in Eq. (5) gets computed for the function of intensity as the constant function and exponential decay function are integrable.

Here, point to remember is that β is not dependent on the dynamics part, and gets derived from solution post Laplacian smoothing application.

$$\beta_{yv} = \frac{\sum_{n=1}^{N} \mathbb{I}(y_n = y) W_{nv} + 1}{\sum_{n=1}^{N} \sum_{v=1}^{V} \mathbb{I}(y_n = y) W_{nv} + V}$$

For α and μ optimization method (HP Approx.), approximation is taken that the logarithm in Eq. (5) by calculating the log within the summation in Eq. (2).

$$\begin{aligned} \mu_y &= \frac{\sum_{n=1}^{N} \mathbb{I}(y_n = y)}{T|D|} \\ \alpha_{ij} &= \frac{\sum_{n=1}^{N} \sum_{l=1}^{n} \mathbb{I}(m_l = m_n) \mathbb{I}(y_l = i) \mathbb{I}(y_n = j)}{\sum_{k=1}^{N} \mathbb{I}(y_k = i) K(T - t_k)} \end{aligned}$$

In the above equation $K(T - t_k) = 1 - exp(-\omega(T - t_k))$ gets calculated by integrating $\kappa(t - t_k)$. In an alternate approach, it is found that parameters using joint gradient-based optimization over μ and α, using derivatives of log-likelihood $\frac{dl}{d\mu}$ and $\frac{dl}{d\alpha}$. In optimization, the process operates in the logspace of the parameters in order to ensure positivity, and employ L-BFGS approach to gradient search. Moreover, it initializes parameters with those found by the HP Approx. method. It is used to normalize the decay parameter ω, in our case to 0.1.

- Transformer Point Process Modeling:
  Transformer point process modeling [3] is an attention centric pipeline that is broadly applied to translation modeling applications and NLP.

  The transformer model can be interpreted with the help of a series of examples, assuming an event sequence, $\mathcal{S} = \left\{\left(t_j, k_j\right)\right\}_{j=1}^{P}$ comprises P events, and every event occurrence belongs to type $pj \in \{1,2,\ldots,P\}$, with aggregated occurrences of P types. Every pair (t_j, p_j) belongs to an p_j type of event, which take place at t_j time.

  The ground hypothesis of transformer-based point process modeling is to attain a self-attention module. It is quite different from RNNs, as the attention module does not follow the recurrent structures. Although the model handles the temporal dynamics of the inputs provided, using time distributed dynamics and batch processing. Hence, similar to the positional encoding methodology the usage of a temporal encoding method is recommended for obtaining better results.

  Figure 2 provides an explanation of how each event sequence S gets connected with the embedding layers, followed by the self-attention modules. The output will be embeddings generated of event S, with knowledge of prior events encoded inside as a knowledge layer.

  The underlying equation summarizes the story behind the process:

$$\left[\mathbf{z}\left(t_j\right)\right]_i = \begin{cases} cos\left(t_j/10000^{\frac{i-1}{M}}\right), & \text{if } i \text{ is odd,} \\ sin\left(t_j/10000^{\frac{i}{M}}\right), & \text{if } i \text{ is even.} \end{cases} \tag{6}$$

  The model utilizes trigonometric equations for building a temporal encoding layer at a granularity of every time stamp obtained. At every t_j, the module deterministically generates $\mathbf{z}\left(t_j\right) \in \mathbb{R}^M$, where M is the encoding dimension. It has

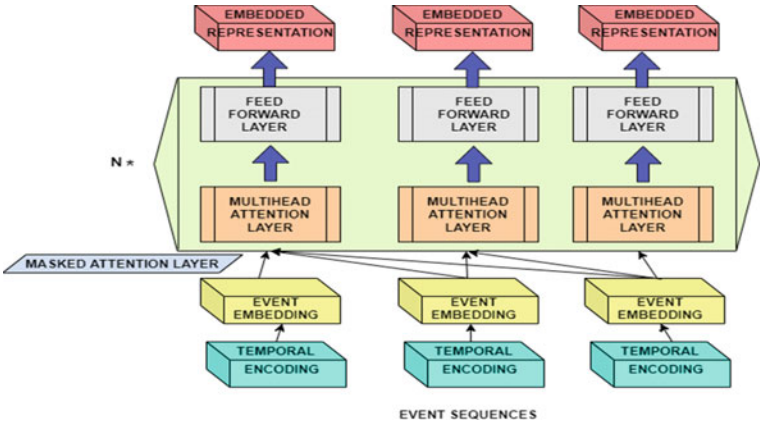


Fig. 2 Transformer Point Process Model

the feasibility of employing several encoding methods, like the position relativity representation module, where instead of predefining, two temporal encoding matrices are generated. The model has the feasibility of training the matrix of embedding $U \in \mathbb{R}^{M*R}$ for the type of events, where the r^{-}th column of U represents an embedding of dimension M for r type of event. Considering an event r_j, with the one-hot encoded (a vector space of r-dimensions with 1 s for the r_j-th point, and the rest of the indexes contains 0 s), embedding as Ur_j. For an event with the relative time stamp (t_j,r_j), the event collation Ur_j and the temporal embedding $z(t_j)$ resides in RM. The embedded representation of the event sequence $\mathcal{S} = \{(t_j, r_j)\}_{j=1}^{L}$ can be specified as follows:

$$X = (UY + Z)^T \tag{7}$$

In the equation [4] mentioned above $\mathbf{Y} = [\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_L] \in \mathbb{R}^{K \times L}$ is the collective event types, their encodings, $\mathbf{Z} = [\mathbf{z}(t_1), \mathbf{z}(t_2), \ldots, \mathbf{z}(t_L)] \in \mathbb{R}^{M \times L}$ and is the cumulative encodings of different event types. Here the key insight is $X \in \mathbb{R}^{L*M}$ and the entries of X belong to the embedded event in order of their relativity. Next X will be passed through the self-attention layer, and output S is formulated as-

$$\mathbf{S} = \text{Softmax}\left(\frac{\text{QK}^\top}{\sqrt{M_K}}\right)\mathbf{V}, \text{ where } \mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \mathbf{K} = \mathbf{X}\mathbf{W}^K, \mathbf{V} = \mathbf{X}\mathbf{W}^V \tag{8}$$

In Eq. 9, {K},{Q}, and {V} represent the key, query, and value matrices generated from various transformations of X, and $\mathbf{W}^Q, \mathbf{W}^K \in \mathbb{R}^{M \times M_K}, \mathbf{W}^V \in \mathbb{R}^{M \times M_V}$ are the coefficients of relative importance of numerous linear transformations, performed in order of their relativity. The usage of multi-head self-attention is recommended for the reduction of input feature space. Various outputs from the attention layer S_1, S_2, …, S_H are generated using weights $\left\{\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V\right\}_{h=1}^{H}$. The final outcome can be generated as follows:

$$\text{S} = [\text{S}_1,\ \text{S}_2,\ \text{S}_3,\ \ldots..,\ \text{S}_\text{H}]\text{W}^O \tag{9}$$

In Eq. 10, $\mathbf{W}^O \in \mathbb{R}^{HM_V \times M}$ is the accumulation matrix. The self-attention module identifies the events having their occurrence, at a certain distance from the present time space. Column j, in the attention layer weights Softmax $\left(\mathbf{QK}^\top/\sqrt{M_K}\right)$ denotes the dependency of event t_j on historical event occurrences, whereas Recurrent neural network models encode history information in relative sequence, i.e., t_j state has dependencies on t_{j-1}, which eventually has a dependency on the t_{j-2} state and so on.

Outputs of the attention layer is next propagated through a feed-forward neural network architecture, where it generates encrypted representations q(t), provided as input to the layer:

$$\mathbf{H} = \mathrm{ReLU}\left(\mathbf{S}\mathbf{W}_1^{\mathrm{FC}} + \mathbf{b}_1\right)\mathbf{W}_2^{\mathrm{FC}} + \mathbf{b}_2, \mathbf{q}\left(t_j\right) = \mathbf{H}(j,:) \tag{10}$$

In Eq. 11, $\mathbf{W}_1^{\mathrm{FC}} \in \mathbb{R}^{M \times M_H}$, $\mathbf{W}_2^{\mathrm{FC}} \in \mathbb{R}^{M_H \times M}$, $\mathbf{b}_1 \in \mathbb{R}^{M_H}$, and $\mathbf{b}_2 \in \mathbb{R}^M$ represents the hyper parameters constituting the neural network, and WFC 2 has similar set of columns. The resultant metric H $\in \mathbb{R}^{LXM}$ provides embedding's of the different event occurrences in the input space, where each entry belongs to an individual event. While the computation of the output from the attention layer S(j,:) (the j-th row of S), look forward positions have been masked. It will be helpful in avoiding the dependency on look forward events. Figure 2, illustrates the architecture of Transformer Hawkes Model.

- Continuous Time Conditioning Intensity:
  Dynamics of event occurrences is formulated as a continuous time conditioning intensity function. Equation 10 produces the embeddings for different time stamps, and the intensity associated also follows the discrete distribution. Considering $\lambda(t|H_t)$ as the model's conditional intensity metric, where $H_t = \{(t_j,k_j): t_j < t\}$ represents the historical events up to time t, it can be derived from the various intensity functions corresponding to specific event types. Every k $\in$ {1,2,…,K}, formulates $\lambda_k(t|H_t)$ to be the conditional probability density function, for event occurrences of type k. and can be defined as follows:

$$\lambda(t \mid \mathcal{H}_t) = \sum_{k=1}^{K} \lambda_k(t \mid \mathcal{H}_t)$$

Here the intensity function of a specific form can be represented as:

$$\lambda_k\left(t \mid \mathcal{H}_t\right) = f_k\left(\underbrace{\alpha_k \frac{t - t_j}{t_j}}_{\text{current}} + \underbrace{\mathbf{w}_k^{\top}\mathbf{h}\left(t_j\right)}_{\text{history}} + \underbrace{b_k}_{\text{base}}\right) \tag{11}$$

Here, the time interval $t \in \left[t_j, t_{j+1}\right)$, and $f_k(x) = \beta_k \log(1 + \exp(x/\beta_k))$ represents the softplus function with "softness" hyper parameter as β_k. The selection of the function can be explained with twofold justification: the softplus function provides the assurance of positive intensity; and the "softness" of the function ensures the stability of computation, avoiding the chances of dramatic fluctuations in the intensity.

- The "current" influence [5] represents the degree of influence between multiple time points t_j and t_{j+1}, and α_k signifies the relative importance associated with them. If $t = t_j$, then a new observation arrives in, and its influence will be 0. If $t \to t_{j+1}$, the conditional intensity function loses its nature of continuity.
- The "history" term comprises dual significance [6]: the W_k vector is responsible for transformations in the embedded states of the Transformer Point Process

Pipeline converts to a scalar entity, and the embedding *h(t)* encodes past events till time space t.

The "baseline" intensity signifies the likelihood of an event incident without taking into consideration the dynamics of historical event occurrences.

4.2 Label Propagation Algorithm

Label propagation was originally introduced by Xiaojin Zhu and Zoubin Ghahramani in the year 2002. This is a semi-supervised algorithm which will be helping the needy population to locate both the financial and medical resources nearest to them without any hassle. The algorithm behind the framework is as follows.

The framework forms a graph architecture [7] with the resource locating feeds as nodes. Next it puts a weighted edge between the nodes using time-centric distance calculation and edge detection algorithms. The framework utilizes dynamic time wrapping-based algorithms for achieving calculating the similarity between the journeys. We will be explaining the process with an example. Assuming we have a sample dataset (2D) that consists of only a few samples, as presented in Fig. 3. There exists binary classified feeds and one among them is un-labeled. The samples represent different nodes of the graph. The framework will be connecting each node with any other node based on dynamic time wrapping-based similarity calculation, and annotate the edges with the resultant distances (Fig. 4).

The algorithm will assign higher weights to the nodes lying closer to each other through the Gaussian transformation using the radial basis function [8].

$$w\left(x, x'\right) = e^{-\left(\frac{x-x'}{\sigma}\right)^2} \tag{12}$$

Here x and x' are the resources located in the surrounding space. If the consumers and the resources are really close, i.e., $|x\text{-}x'|$ is close to 0, their assigned weightage [9]

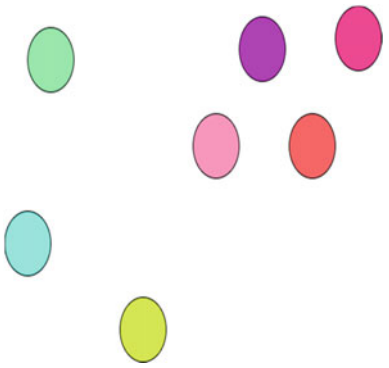


Fig. 3 Data points in n dimensional space

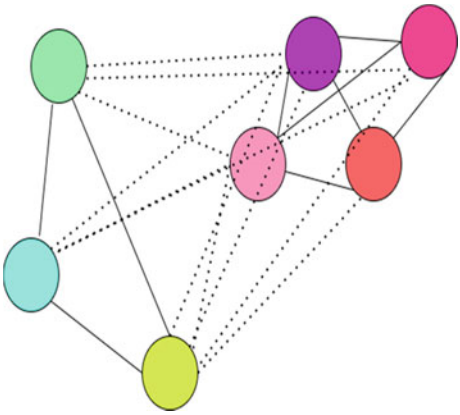


Fig. 4 Label Propagation based on similarity

will be approaching 1 and in case they are further apart from each other, the weight will tend to approach 0.

σ is a hyperparameter to be optimized over time based on model performances and loss reduction in allocation efficiency.

Next for locating the nearest resource existing from an individual, the framework will start a random walk originating from its locus. One step of the walk refers to the movement from a particular node to another one except the source node itself. Edges associated with a higher weightage will be prioritized for optimal efficiency. It applies Markov chain modeling along with K nearest neighbors-based algorithms for setting up the priorities between the crosswalk selections. Markov chains are a stochastic modeling method that explains the occurrence of a series of possible events where the likelihood of an individual event depends only on the state attained from the previous event maintaining the sequence history for data-centric decision building. Finally, the label of each of the points will be assigned to the nearest neighboring resources.

4.3 Framework Modeling

A point process model is a collection of randomly located points located on a line or Euclidean space. The modeling algorithm is a stochastic process underlying the occurrences of events in a particular time/space. The occurrences of every single event including different pandemics like Covid-19 are very likely to follow a particular trend, and our intention will be to capture and analyze such patterns at the utmost granular level, so that those learnings can be utilized in the future to prevent or fight against such pandemics. For example likelihood of a person getting a severe heart attack increases when the person has high blood pressure and belongs to the elderly age groups than in his middle ages with stable blood pressure. The panic created due to instability of sales and revenue generation of a corporate firm in one country

can cause similar events in other countries, which will further lead to a disaster in the public valuation of the company. Taking an example of the Amazon forest that suffered from wild file this year, created awareness in the entire world which will eventually reduce the likelihood of another forest, facing wildfire going forward. These examples helped us understand the fact that the likelihood of the occurrence of an event can be increased or decreased following the patterns in the sequence of prior events happening around the world.

If the events happened to increase the likelihood of the occurrence of the future event, those events are known as stochastically [10] excited or self-excited events. The example of a person having a heart attack falls under this category, whereas if the likelihood of a similar event is decreased like the example of the impact of company valuation, then it is known as stochastically inhibited or self-regulating events (Fig. 5).

The proposed solution intends to capture and analyze these patterns present in the surveillance feeds, collected from the governing authorities. Once the patterns are identified they can be clustered based on comprehensive dynamic time wrapping distance measurement. The proposed framework [11] intends to capture all possible sources of information, granulated at minimum source level to capture live information feeds capturing demographical, hospitality, and mobility metrics. The model will identify the chaining pattern existing between the events, proactively forecast the future trend before time and assist governing authorities with alarming alerts, in case of a requirement for additional care. The framework is capable of sending out alarming alerts via different mediums. This pipeline of processes will eventually help in resource optimization and strategic execution. The forecasting layer also outputs regular monitoring status checks along with forecasted status predictions for the next 90 days, enabling greater visibility and control over the pandemic (Fig. 6).

During the catastrophic pandemic, there has been a severe crisis in employment opportunities for the working class. Although humanitarian aids and supporting offers

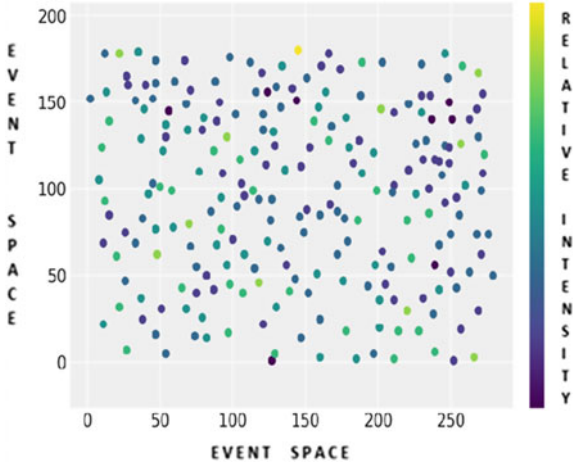


Fig. 5 Relative Intensities among the events

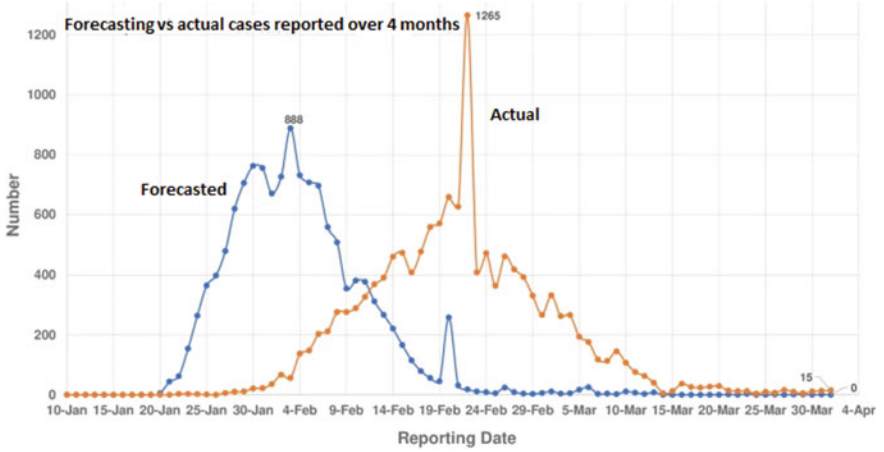


Fig. 6 Actual vs Forecasting

from government agencies, non-profitable enterprises were offered from every corner of the country to meet their most-pressing needs, connecting the help providers with the people in need at scale seems to be the biggest challenge. The proposed solution (U/R Mapping layer) [12] will enable the population to locate the resources/utilities closest to them at their fingertip using a semi-supervised learning algorithm. The algorithm is inspired by a technique from experimental psychology called spreading activation networks. The algorithm employs the points, present in the dataset to be connected in a graph-based network on their relative distances in the input space. The weight matrix of the graph is then normalized symmetrically, much like spectral clustering methods. Information is passed through the graph, which is adapted to capture the structure in the input space. Finally, the label of each un-labeled point is set to be the class from which it has received the most information during the iteration process (Fig. 7).

The surveillance layer is responsible for tracking the performance of model prediction and tries to learn from the mistakes, done in the past. The base idea behind this is to make the pipeline learn from the environment by interacting with it and receiving rewards for performing correct actions. Interaction with the environment stands for gathering the event occurrence pattern over a period of time and analyzing the hidden trends present in the data.

$$\text{Reward function} = \frac{(0.75 * \text{ previous batch count}) + (0.85 * \text{ current batch count})}{(0.75 * \text{ previous batch size}) + (0.85 * \text{ current batch size})} \tag{13}$$

The algorithm will refine with time using incremental learning based exponential smoothing based reward function mentioned in Eq. 13. The key insights derived will be backpropagated to the modeling algorithms to penalize the outdated patterns and

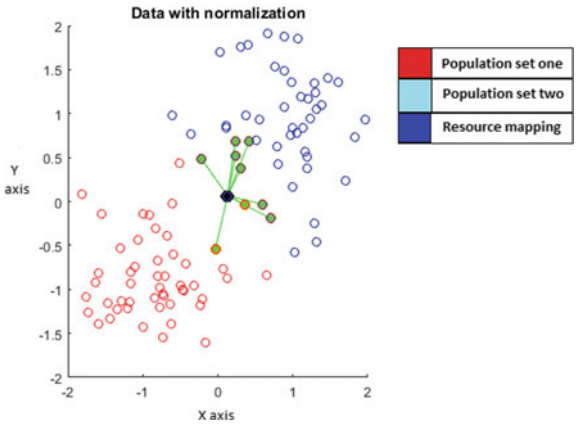


Fig. 7 Label Propagation

promote the recent trends to penalize the outdated patterns and promote the recent trends.

5 Conclusion

In this book chapter, we have introduced an Artificial Intelligence based framework to tackle pandemics like Covid-19. We have applied point process modeling using the Hawkes process [13] for the next Covid hotspot prediction, which will help the governing authorities to receive prior alerts and be prepared for the upcoming fights against these deadly pandemics and save humanity. This pandemic has ripped the world apart in every possible way from academic to financial, physical to physiological. The resource/utility mapping layer of the framework intends to solve the problem of employability and nearest food and medical resource allocation using dynamic time wrapping Label propagation methods. This pandemic has triggered one of the worst job crises in the entire world and it is being estimated that it will increase the poverty rate along with widening inequalities in the years to come. The governing authorities need to brainstorm collectively to take every possible action to stop these crises from turning into a serious social crisis. Our framework attempts to contribute to the same cause. It has been made capable of learning incrementally over time to adapt to the new environment and update itself for providing the best possible assistance toward re-establishing the normal.

References

1. Bacry, E., Mastromatteo, I., Muzy, J.-F.: Hawkes processes in finance. Market Microstruct. Liq. **1**, 1550005 (2015)

2. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**, 157–166 (1994)
3. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018). arXiv preprint. arXiv:1810.04805
4. Hawkes, A.G.: Spectra of some self-exciting and mutually exciting point processes. Biometrika **58**, 83–90 (1971)
5. Li, S., Xiao, S., Zhu, S., Du, N., Xie, Y. and Song, L.: Learning temporal point processes via reinforcement learning. In: Advances in Neural Information Processing Systems (2018)
6. Linderman, S., Adams, R.: Discovering latent network structure in point process data. In: International Conference on Machine Learning (2014)
7. Mei, H., Eisner, J.M.: The neural hawkes process: A neurally self-modulating multivariate point process. In: Advances in Neural Information Processing Systems (2017)
8. Xiao, S., Yan, J., Yang, X., Zha, H., Chu, S. M.: Modeling the intensity function of point process via recurrent neural networks. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
9. Yang, S.-H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., Zha, H.: Like like alike: joint friendship and interest propagation in social networks. In: Proceedings of the 20th International Conference on World Wide Web (2011)
10. Yin, W., Kann, K., Yu, M., Schutze, H.: Comparative study of cnn and rnn for natural language processing (2017). arXiv preprint. arXiv:1702.01923
11. Zhang, Q., Lipani, A., Kirnap, O., Yilmaz, E.: Self-attentive hawkes processes (2019). arXiv preprint. arXiv:1907.07561
12. Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., Leskovec, J.: Seismic: A selfexciting point process model for predicting tweet popularity. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2015)
13. Zhou, K., Zha, H., Song, L.: Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In: Artificial Intelligence and Statistics (2013)

Role of Data Science in Programmatic Advertising

Aditya Jain, Kashyap Barua, and Monali Barbate

Abstract The ability to capture the attention of consumers in the online ecosystem has become ever more daunting. Millions of websites, digital properties, and web applications are being viewed daily and this is where the need to select the best platform to show the right advertisement and content to the correct cohort of users becomes challenging, with the diversity. This is where programmatic advertising comes into play, to be able to solve these problems. Programmatic advertising is the buying and selling of online advertising and this piece of technology makes the automation seamless for transactions. Numerous programmatic platforms have made it convenient to access any format and channel programmatically, which includes mobile, desktop, laptop, audio, and connected TV. Moreover, segmentation and targeting techniques are implemented which makes sending the correct ad to the right cohort of users more efficient making the ad buying and selling experience more economical than ever. Algorithms and programs make the process of buying and placing ads very optimal, thereby removing the manual processes within these steps. With Data Science coming into the ad-space domain, how companies can control their advertising costs has become unparalleled with the traditional methods. Real-time bidding, dynamic budget allocation, and contextual targeting are to name a few when it comes to the applications of data science in the programmatic advertising domain. This chapter would primarily focus on the digital advertising domain and the role of data science in solving some of the complex problems in the same. Through this chapter, we would try to understand the basics of Digital Advertising along with the data science's association with the same, and then deep dive into some of the popular terminologies and techniques in the advertising world. Data science has a lot to offer in the programmatic advertising world and we want to showcase some of those examples here, which helps our readers understand the applications perspective of Data Science in the modern world. Certain factors and advantages come into play when data science comes into the advertising area along with its

A. Jain · K. Barua (✉) · M. Barbate
MiQ Digital India, Bengaluru, India
e-mail: kashyapbarua@gmail.com

A. Jain
e-mail: aditya@miqdigital.com

S. S. Rautaray et al. (eds.), *Data Science in Societal Applications*, Studies in Big Data 114, https://doi.org/10.1007/978-981-19-5154-1_3

applications. We will also discuss some of these advantages in depth as to how advertisers and brands benefit from these out-of-the-box solutions available to them. The availability of Petabytes of advertising data made accessible to programmatic media partners has allowed them to leverage data for these advertising decisions through supervised learning techniques, allowing them even more optimized advertising strategies for their clients. The modern era is a transition toward programmatic and data science for the advertising world. It has become imperative to understand these nuances and applications to help netizens and other researchers stay updated with these technological progressions.

Keywords Digital advertising · Programmatic advertising · Real-time bidding

1 Introduction

People spend more time on the Internet than ever before and the need to go digital has become very important. Any business, irrespective of its size, should have an online-first approach to reach the users. According to a study, 64% of people are influenced by advertisements online before they go ahead to buy that product. Digital advertising is the process of publishing promotions online on various platforms like social media, websites, gaming apps, and many other applications that can be accessed digitally by the people. With a vast majority of the population having access to the Internet, that is, where the actual action has transitioned toward. According to Google, 3.5 billion searches by people in a day and this value increases by 10% year over year. With the right approach to digital advertising, brands and advertising partners can boost their reach to audiences along with maximizing their sales.

From the traditional process of publishing ads on creatives across different websites through a brand-publisher relationship, technology has led to a modern approach to advertising, i.e., programmatic advertising [1]. Since the time when the first ad banner was run by AT&T in 1994 (which was placed on http://www.hotwired.com), a lot has evolved. From the static form, advertisement banners have become interactive which support rich content along with variations in sizes for different types of display devices.

As the supply chain grew with the massive digital adoption, there emerged an unbalance in the chain wherein there were more suppliers than buyers, which led to unused supply inventory left unsold for any brands [2]. Along with that, certain other complexities spewed like variability in ad content type, genres of websites, the advertising network, and many others. Before programmatic, advertising was as simple as a one-to-one relationship between the advertiser and the publisher. This approach leverages technology to buy and sell ad inventory through an automated and data-driven approach across various channels like display, video, audio, etc. [3]. It also focuses on individual impressions based on the target audience properties like geolocation, daytime, language as well as other demographic and behavior data. Programmatic advertising uses a data-first approach to simplify the complexities in

Fig. 1 Programmatic supply chain

the modern advertising domain, which not only helps advertisers target the right audience, but also optimizes various cost metrics like CPC (Cost-Per-Click), CPA (Cost-Per-Acquisition), and CPM (Cost-Per-Mille).

In a typical programmatic chain, Fig. 1. demonstrates the business process flow from the advertiser side to the audience who view the advertisements.

- Advertiser—Brands or companies who promote their products or services to the consumers through digital media
- DSP or Demand Side Platform—Allows advertisers and agencies to buy ad spaces to showcase their promotions
- Ad Exchange and Networks—The ad network aggregates the supply side publisher spaces to make it available to the demand side (advertisers) as per relevancy. Ad Exchanges facilitate the buying and selling of media advertising inventory, the prices of which are determined by bidding from multiple ad networks
- SSP or Supply Side Platform—These platforms allow the suppliers to connect their inventory to the ad exchanges and DSPs
- Publisher—They are the creators of ad spaces across the internet through websites and blogs
- Audience—These are the users who consume the advertisements and digital media through the publishers online.

2 Overview of the Advertising Industry

Before we go into the depths of how the programmatic industry works along with the practical applications of data science in the same, we want to touch up on some of the important terminologies of the advertising industry. The digital advertising universe is a vast sphere, and the business has been gaining traction for the past couple of years. The core idea behind this industry is the interaction between the advertiser and the publishers, wherein the advertisers want the most reach for their brand with the least amount of spend whereas publishers want their share of payments for promoting these brands on their websites [4]. There are many metrics that this line of business considers while gaging between a right advertising strategy with a wrong one, below are some of them:

- Impressions, these are the number of times an ad banner or creative viewability by a user. For example, in a google search, if a user sees an ad for a product on Amazon three times, the total impression for that would be 3.
- Clicks are the count of the number of times the user/users click on the ad creative on a website or blog.
- Click-Through-Rate or CTR is an important metric that helps advertisers and agencies understand the number of clicks out of a total number of times the ad creative was viewed. Let's say that a CTA (Call-to-Action) was available on a blog pot and 900 users saw the same out of which only 45 clicked on it, the CTR is calculated to be 5%.
- Cost-Per-Click or CPC is the cost associated with the number of times an ad is clicked by viewers. This metric is not dependent on the fact about how many times the ad was viewed and only charges based on clicks.
- Cost-Per-Mille (CPM) is the cost metric that is dependent on the impressions for any ad, i.e., this is the cost paid for every 1000 impressions for that ad.
- Conversion Rate has been used by advertisers and agencies to understand the performance of the desired action that an ad intends for the user to take. For example, if an ad wants users to sign up for an upcoming webinar, the number of users registering for the same is going to determine between a good conversion rate and a bad one.

Campaign is a strategic plan for achieving a clear and collective goal for a brand or organization. When these entities want to reach the users about their ideas and promotions through digital media, they often lean on creating campaigns based on these goals. Now, when it comes to running campaigns on different platforms there are certain types of campaigns that are used a lot [5]. To elaborate on some of these, they are listed below:

- Branding Campaigns: These types of campaigns are often based on strategic efforts to raise awareness, develop or even change a brand while on the contrary, a marketing campaign is the promotion of a product or service for the increase in sales
- Retargeting Campaigns: Such campaigns are very prominent in the digital landscape, wherein visitors are reminded of the products or services that they have viewed in the past without converting. This is quite prominent in e-commerce websites like Amazon wherein if a user views some products without purchasing them, they are shown the same product on ad spaces on other websites and blogs
- Contextual-Targeting Campaigns: In this type of display advertisement, the ads are placed on websites in such a way that they are directly relevant to the product being sold in the ad. Contextual ads leverage user session data (using browser cookies) to understand a user's intention and interests toward a product. This type of campaign ensures relevancy for the users and in turn leads to more conversions for the brand
- Geofence-Targeting Campaigns: This is a type of location-based targeting wherein users are targeted with ads by using their smartphone location. If a user is about to enter a shop or is just leaving the same, then the ads pertaining to products

from that store and their products are shown to the user. The level of relevancy is unsurmountable with this type of campaign practice, since the modern society is driven by smartphones and brands can have greater accessibility to these devices for advertisement as per their requirements [6].

Having discussed some of the key metrics used in the advertising industry, this paves the way to getting into the depths of the industry along with the use cases of data science in the programmatic advertising journey.

3 RTB—Real-Time Bidding (Based on Game Theory) and Second Price Auction

Programmatic advertising is a huge ecosystem, where big volumes of inventory move at a ginormous pace. To make the existence of this marketplace possible, the technology must be automated for buying media, i.e., ad-space as opposed to traditional (often manual deal hinged buying). Programmatic media buying utilizes data insights and algorithms to serve ads to the right user at the right time, and at the right price [7]. Firstly, programmatic media buying can be categorized into three different types:

- Real-time bidding (RTB): Also known as open auction, RTB is when inventory prices are decided through an auction in real time. As the name suggests, this is open to any advertiser or publisher. RTB is a cost-effective way to buy media with a large audience.
- Private Marketplace (PMP): These are like open auctions, but PMPs have restrictions on who can participate. Only selected advertisers have access to PMPs on an invite-only basis. However, in some cases publishers may have a selection process that allows advertisers to apply for an invitation.
- Programmatic Direct: This is when a publisher bypasses auctions, selling media inventory at a fixed cost per mile (CPM) to an advertiser (or multiple advertisers).

Programmatic media is increasingly skewed toward real-time buying, real-time bidding. Over time the RTB ecosystem has evolved over multiple pricing models in view of profits for the participating parties. Previously, ads were served through a process called waterfalling, where the bid of each advertiser would need to meet the minimum price set by the publisher. The programmatic system has shifted to header bidding auctions, in which the highest bidder wins.

Waterfalling is a method in which the publisher sets the priority for each advertiser or ad network they are connected to. They specify a minimum acceptable price for this ad placement. The inventory slot is then sequentially offered to demand partners in order of priority. Once someone meets the price floor, the impression is sold. This method is slow; however, it drives a high fill rate for the ad slot. The inventory is often going to the one who is first in line, not the one who bids the highest. Now we have RTB and header bidding as a product of its evolution.

Header Bidding is a method where Publishers exhibit their inventory at several ad exchanges [8], which in turn allow multiple demand partners to bid on those “lots” simultaneously. In header bidding, the highest bid wins.

RTB versus header bidding in programmatic real-time bidding and header bidding are often used as synonyms today, however, there is a difference. A modern RTB often involves header bidding technology, so both the RTB and header bidding concepts may be applied to explain what’s happening inside your DSP. RTB has a wider definition. If header bidding is not implemented by the publisher, the programmatic platform will use an older ad tag model to sell inventory. So, an RTB protocol can be based on both ad tags, header bidding, or even an SDK depending on the technology stack implemented by the publisher.

First-price versus second-price RTB in programmatic header bidding auctions in programmatic can be first-price and second price. A second-price auction is the initial model of the RTB process, where the highest bidder doesn’t pay the price equivalent to their bid, but $0.01 higher than the bid of the second bidder.

A first-price header bidding auction is more beneficial for publishers, as the highest bidder pays exactly the price equivalent to their bid. It is the simpler framework that allows maximizing yield from each inventory sale.

4 Data Available in Programmatic Advertising

Before we discuss how data science is applied in digital advertising, let’s first have a look at the data that is available. Different organizations that are placed differently in the advertising landscape have access to different types of data. The exchanges have access to granular data on the availability of impressions, and their price. They are, however, less concerned with the performance of the said inventory. The advertisers on the other hand often do not have access to this data. They are limited to the data of their campaigns and the impressions that they have won. For the purpose of this text, let’s consider datasets that are available in the public domain. One such dataset is iPinYou dataset. This dataset contains real-time bidding log, conversion log, and click log.

Some of the important fields in the bidding log are as follows:

Bid ID: Unique identifier of this auction.

iPinYou ID: Hashed unique identifier of the user.

Ad Exchange: DSPs usually place bids via multiple ad exchanges. This column indicates the ad exchange of the current auction.

URL: URL of the page where this ad impression is available.

Region ID: Geographic Region of the current user.

IP Address: IP Address of the current user.

Ad Slot ID: Unique identifier of the Ad slot. Pages can have multiple ad slots.

Ad Slot Visibility: This is also referred to as the "Fold". This indicates whether the ad slot is above the fold or below the fold.

Creative ID: The unique identifier of the ad creative for which the bid is sent.

Advertiser ID: Unique identifier of the advertiser for whom the current bid is placed.

Ad Slot Height: Height of the Ad slot.

Ad Slot Width: Width of the Ad slot.

User Profile IDs: Profile IDs associated with the current user.

Apart from this, there are multiple fields providing information about the user's category, URL category, etc. The bidding log provides information before, an auction happens. This is the information based on which advertisers will decide on:

- Whether to show an advertisement on this inventory
- Which of the many advertisements to show on this inventory?
- What is the viable price for this inventory?

Once a decision is made on this, and a bid is sent to the exchange, further data is funneled to the impression log. This log also contains information on clicks and converts. Let us have a look at the important fields of this dataset. The following list only includes additional fields with respect to the previous list.

Bid ID: Unique identifier of the auction that led to this event.

Paying Price: Actual price that is billed toward this impression.

Log Type: Whether this row of data represents an impression, click, or conversion.

If you look closely, you will see that the features of the bidding log, and impression click, and conversion log can be classified into two categories:

- Contextual Features: These are the features that answer the question "Where is this impression being shown?" Ad Exchange, Ad Slot ID, Ad visibility, etc. features fall under this category. Location-related features also fall under this category.
- Behavioral Features: These are the features that convey the information about user's past behavior like websites they have visited and their interests. The iPinYou ID field (Cookie ID) along with the User Profile IDs field convey this information.

Due to recent advancements, configurable fields are also available in some of these datasets. These fields are captured from special pixels often referred to as "advance pixels". These advanced pixels allow the advertiser to associate many (often in the range of 50) key-value pairs to each event in the conversion log. These key-value pairs usually include information about the purchase price of the product, advertiser specific user segment, etc. This allows a superior approach to modeling where each convert is not treated equally and is considered in the light of data available from the advance pixels. For example, instead of modeling for a conversion ratio, we can build a model that prioritizes high spending instead of simple conversion.

5 Structuring Programmatic Advertising as a Data Science Problem

All programmatic advertising revolves around answering the following questions:

- What do we want to achieve from the digital campaign?
- Who should we show an advertisement?
- Where should we show an advertisement?
- Which advertisement should we show?
- Price at which we want to show an advertisement.

The first question "What" is the most important question that needs to be answered first, before answering the next four questions.

In this section, we will see methods that are used to implement the answers to questions 2–5.

Who

The answer to this question defines the audience of an advertisement campaign. The two of the most common methods to address this are as follows:

- Lookalike segment
- Interest-based segments.

The objective of lookalike segmentation is to create a list of Cookie IDs that have a distribution like a known set of Cookie IDs. This is commonly used to reach people that are expected to be similar in behavior to ones already visiting an advertiser's website. In case of lookalike segments, the advertiser is expected to have a set of Cookie IDs that have already performed a desired action like purchasing a product or visiting a webpage.

One way to solve this is by clustering. Let S be a set of users we want to use as seed. Let H_i be the history of previous actions taken by a user i in S.

Let A be the set of all users. Let Ji be the history of previous actions taken by a user i in A. Then if Hi~Jk then we can say that the users are similar.

Once clustering is performed on set J, we can compare it with set H. Clusters closest to elements of set H can be then used as a lookalike audience.

Similar to the lookalike segment, interest-based segments are also lookalike but not to an advertiser's provided segment but to a set of users who are known to have an interest.

With the advent of GDPR, Cookie ID-based segments are losing their shine due to privacy concerns.

Where

Often where the advertisement was shown impacts the response that could be expected from it. The 'where' question answers which:

- URL to show and ad on
- Which Geography to show an ad on?

- What kind of device should be targeted?
- What time of day should the advertisement be shown?

This is a binary classification problem where the aim is to predict the probability of an action/response on a given set of contextual features.

Let us take an example of click prediction. For a given set of context combinations, we know how many users were shown an advertisement, and how many of them ended up clicking on the advertisement.

Let context C be the set of features {URL, time, device type} where a particular impression was shown. Let y be the target variable that represents a click. If an impression leads to a click, the value of y will be 1, 0 otherwise.

Then our goal is to train a model to predict $P(y = 1|C)$.

In practice, this is modeled using deep learning or gradient boosted trees.

Which

Typically, advertisers run hundreds of different creatives. Not all creatives perform similarly. It is important to prioritize creatives that have a high likelihood of encouraging the user to take the desired action.

Multi-armed bandits is a family of algorithms that allow for maximization of return when there are competing choices with unknown performance.

This problem in digital advertising is called dynamic creative optimization.

Price

Second price auction which is utilized extensively in RTB encourages value-based bidding [9]. Participants are encouraged to set the bid equal to the value they are deriving from the ad impression. Let us take an example: Let the cost per click goal be $2 Let the click through rate be 0.1%

This would mean that for every 1000 impressions, we can expect 1 click. Going by this number, and with our goal of $2 CPC, we can bid a maximum of $2 per 1000 impressions.

In practice, however, bidding by value leads to overbidding.

Let us consider a simple example of a campaign having $10 to spend. Summary of campaign setup:

Bid Price: $5.

Budget: $10.

Once the campaign goes live, it will start participating in RTB auctions. Consider the following order of auction requests for a single context:

Order of auction	1	2	3	4	5	6	7	8	9	10
Win Price	1	1	4	4	1	2	1	1	1	2

For simplicity, let us assume that the win price is per impression and not per 1000 impressions.

Given this order of auctions received, we will consume all our budget by the time we reach auction 4 since our bid price is $5. This will result in an average cost of $2.5 = 10/4.

However, if the bid price would have been $2, we would have lost auctions 3, and 4. However, we would have won auctions 1,2,5–10 while spending the same $10 bringing the average cost down to 10/8 = $1.25.

Consider another type of campaign with an impression budget:

Impression Goal: 5 impressions.

Bid Price: $5.

Using the above table, we will win auctions 1–5, with an average cost of $11/5 = $2.2. However, if the bid would have been $2, the average price for 5 impressions would have been $6/5 = $1.2. Moreover, if the bid price would have been $1, the average cost would have come down to $1 and total cost would have come down to $5 from $11.

Therefore, knowing how many impressions can be won at any given price for a context helps us bid correctly, in line with our delivery target.

This is often modeled as a regression model that predicts winning probability at any given price point.

6 Application of Programmatic Advertising to Solve Industry Challenges

One of the most intuitive approaches to learning about something is its practical implications and how the technology has been used in the industry to solve real-world problems. In this section, we pick up some industries that have relied on programmatic advertising to tailor their customer's ad experience in the most effective way possible. From airline companies to coffeeshops, from toy stores to bake houses, these industries have been able to unlock the key to the different advertising strategies and data available.

Airline Industry

Airline companies have a very important goal set, i.e., build a brand to drive ticket sales which would allow them to be profitable. This is about an Airline company that used the latest technologies of advertising to segment their users and showcase personalized ads.

The company created cohorts of their customers based on numerous traits like loyalty, whether they bought any tickets with them in the past, or have been frequent flyers with the airline over the years. These factors helped the company understand various factors that have an influence on these user cohorts as to what ads they show them. Based on these insights and data-driven approaches, the company was able to target the users with relevant ads based on their traits. For example, a loyal customer was shown offers for the route that they frequently traveled on which increased the

chances of conversion for the company. The end results demonstrated a 30X return on ad spends which is something that these companies really enjoy, given the amount of money that they invest into their advertising campaigns.

E-Commerce Industry

Online retail companies enjoy targeting users with ads based on their search patterns on their websites. Popular E-Commerce giants quite frequently target ads to their un-converted sect of users based on their browsing habits. So, let's say customer A searched for some books on the website but didn't necessarily convert (make a purchase at the end) while customer B on the other hand purchased a dining table on the same website. Now, the company has enough data so that customer A is targeted with the same books that he/she searched for on different websites and blogs that they visit, which might lead to a conversion from the ad later. On the other hand, customer B might be recommended some chairs to go with the dining table when he/she visits some other websites or during their journey over the internet. This is how these companies leverage data to make the advertising experience for their customers personalized based on their behavior.

This results in more conversions for the companies and customers enjoy how their preferences are nurtured by these companies to recommend things that would go well in their life.

Coffeeshop Company

A very popular coffee shop brand that is available worldwide is one of the top players in the market when it comes to showing ads to their customers based on their needs. This company knows its customers so well that the customers receive advertisements based on what flavor of coffee they prefer and some emerging items that they might love. These personalized ad experiences have led to tremendous growth for the company in terms of ad spends. They also have a personalized app experience for their customers who use their android and iPhone apps, which leads to more conversions and in turn sales [10]. They are also involved in community-based advertisement targeting wherein they send ads based on some campaigns carried out by this cohort of users in the area. This not only improves engagement with brands but also increases their presence in the market and gives an edge over their competitors.

Online Retail Brand

A men's footwear brand faced tough competition from sports brands like Nike and Adidas, with its counterparts having an advantage through online presence. By running online ad campaigns in the form of videos, it aimed to establish its brand identity and align it with the consumers' mindset. Leveraging lookalike audience builds, and techniques like segmentation of users based on demographics and interests, the ads were planned to run on Facebook. The brand did not initially aim to retarget the cart abandoned and purchasers from the cohort of users, but rather focused on optimizing lookalike audiences prospecting campaigns. With a programmatic approach to advertising, the brand saw a soaring Click-Through-Rate on its website which went as high as 233% in the UK market along with the same going

higher in the German market by 800%. These results impact the overall revenue in the end, which becomes the best return-on-investment for using these advanced forms of media campaigns.

These are some of the examples of how programmatic advertising is shaping the landscape for these businesses and it is only a matter of time before all industries would have to rely on programmatic advertising to get their branding and marketing campaigns done. Programmatic advertising has not only made companies rely on the technology but has led to small-scale startups vouching for the approach to get the fruit at its early stage.

7 Digital Advertising and Fraud

Ad fraud or invalid traffic is the process of showcasing online advertising impressions, conversions, or even clicks in a misrepresented manner to generate revenue. This is quite prevalent in the media tech industry which impacts an advertiser's Return on Investment done in media and hampers the brand's reputation in the longer run. According to reports from Forrester, the advertising industry is projected to lose close to $10.9 billion from frauds in advertising. Major contributors to this fraudulency are click bots and click farms available across a myriad of domains over the internet. There are other intricacies to this problem that we want to touch base on, fraudulent advertising also leads to deviations in the success metrics of the campaigns in terms of costs and other KPIs. One of the typical examples to demonstrate this problem is that an ad campaign might show as ineffective to the marketing team due to high impressions but very low clicks ultimately. These impressions might have been from various techniques like ad stacking, pixel stuffing, etc. which sends the impression that the campaign might have not been effective in conversions and interactions, due to skewed campaign metrics. This leads to these teams dedicating time and cost to make changes to these campaigns when the campaign might have turned out more than effective in its true sense. Another problem in this area is that even though the campaign measurement teams might be very aware of these fraudulent activities, they might refrain from calling these issues out and the organization most probably is inclined toward valuing vanity metrics. This leads the teams report on the same just to be on the good books. While they may look good on reports, the ROI for such media campaigns often turns out detrimental for the company in the longer run.

There are a few ways in which a digital advertisement fraud can occur. Ad fraud defrauds businesses and earns money from their advertising efforts. The ad fraud varies depending on the type of campaign and the means.

- Domain Spoofing: It is the practice of mimicking a valuable website and tricking advertisers into paying more for advertising space on the spoofed website than they should. Click Injection is more commonly known as bot activity. The cyber-criminals use malware to generate clicks on ads to falsely inflate spending on those ads

- Cookie Stuffing: Advertisers with affiliate marketing programs often use cookies to track credits to be given to affiliates. IncCookie stuffing fraudsters apply numerous affiliate tracking cookies to a website visitor's browser at the same time just in case they happen to go to a website with an affiliate program later. The merchant's affiliate marketing program gives credit for the visit to the cookie stuffer. This misattribution can consume ad spend and hurt honest affiliates since the fraudster is stealing credit for leads and sales from them
- Pixel Stuffing: This typically creates a 1×1 pixel area, an advertising display that a person will not be able to see. Pixel stuffing can help display dozens or even hundreds of ads on a single webpage and get credit for impressions. These single-pixel ads do not generate results since viewers will never even realize that they "saw" an ad
- Geo Masking: This involves spoofing the IP addresses of the leads they generate to make them look more valuable so they can charge advertisers more for the "leads" they provide
- Ad Stacking achieves the same results as pixel stuffing. The ads are typically stacked over each other, thus inflating the number of ads displayed, where they aren't seen by the user
- Ad Injection is a fraud strategy where criminals use browser extensions, plugins, and malware to put ads where they should not appear or replace the ads on a website with different ads.

References

1. Jun, W., Weinan, Z., Shuai, Y.: Display Advertising with Real-Time Bidding (RTB) and Behavioural Targeting (2017). arXiv: 1610.03013
2. Muthukrishnan, S.: Ad Exchanges: Research Issues. In: Leonardi, S. (Eds) Internet and Network Economics. WINE 2009. Lecture Notes in Computer Science, vol. 5929. Springer, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10841-9_1
3. Iryna, S.: The Evolution of Advertising: From 1971 to Present Days. Smartyads (2019). https://www.smartyads.com/blog/evolution-of-digital-advertising/
4. Phlywheel: Starbucks' digital marketing strategy: How do they do it? https://www.phlywheel.com/articles/marketing-strategy/starbucks-digital-marketing-strategy-how-do-they-do-it/
5. Sarah, B.: Geofencing Marketing: The New Way to Market Your Business. WebFX (2021). https://www.webfx.com/blog/marketing/geofencing-marketing/
6. AcuityAds Inc.: What Is Programmatic Advertising: Programmatic Advertising Platforms (2021). https://www.acuityads.com/blog/2017/12/15/what-is-programmatic-advertising/
7. Klemperer, P.: Auction theory: a guide to the literature. J. Econ. Surv. 227–286 (1999)
8. Ying, C., Ruofei, Z., Jianchang, M.: Bid landscape forecasting in online Ad exchange marketplace. In: KDD '11: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 265–273 (2011). https://doi.org/10.1145/2020408.2020454

9. Grigas, P., Lobos, A., Wen, Z., Lee, K.-C.: Optimal Bidding, Allocation, and Budget Spending for a Demand-Side Platform with Generic Auctions (2021). https://www.ssrn.com/abstract=3841306. https://doi.org/10.2139/ssrn.3841306
10. The Branded Agency: Branding Campaign Examples. https://www.brandedagency.com/blog/branding-campaign-examples

Social Development Data and Societal Modelling: A Study in Indian Context

Rabi N. Subudhi

Abstract As we progress ahead and measure our growth in every field, science and technology or even societal development, every time, our focus is skewed and mostly myopic on what we achieved and what more to be done and how fast to be done. We seldom study what we really missed and skewed our development is. A retrospection and introspection, many a times, help us making a course correction. A development, or any economic growth, is no development if it is not for all and leaves a section of society. This paper looks at different social development models and important development indices which are in use today globally and analyses social development data in the Indian context. It refers to Mahalanobis D-square statistics and other Indian studies and attempts to suggest newer social development indices, using secondary data. It discusses issues relating to 'social-big-data', as available from secondary sources, published by various agencies. Paper presents a conceptual model for measuring the social wellbeing of individuals and families.

Keywords Social development · Development models · Development indices · HDI · Social data analysis

1 Introduction

As technology advances, gathering and retrieval of a large amount of data are now comparatively much easier. We have so many instances of handling very huge databases in real time, in many fields. Millions of share holders' data in the stock market or thousands of passengers' reservation data, railways or airlines, or the daily transactions of a big retail chain are some such examples, where very huge databases are handled, and used for decision making. Apart from business transaction data or administrative data, we also use large scientific experimental data for our research purpose. This data could be both quantitative as well as qualitative data. Here, in this paper, we shall confine our discussion to social development-related data.

R. N. Subudhi (✉)
School of Management, KIIT University Bhubaneswar, Bhubaneswar, India
e-mail: rabisubudhi@gmail.com

S. S. Rautaray et al. (eds.), *Data Science in Societal Applications*, Studies in Big Data 114, https://doi.org/10.1007/978-981-19-5154-1_4

It has been of enormous research interest to study and analyse the complexity and diversity of different demographic strata of any society or our humanity. How and why a region or a community or a religion or a caste/ creed is not so 'developed', as compared to other strata! A construct like 'social development' or 'social acceptance' is quantified and measured through several 'representative' operational indicators (or variables), differently by different scholars. But the commonality amongst all such known measures, as outlined in next section, is the use of a large amount of socio-economic and demographic data. We may call it social big data. One such example can be given on the record single-day COVID vaccination data (on 18.9.21, coinciding with the birthday of Indian PM, Sri Narendra Modi, when 25 million or 2.5 crore persons were vaccinated in India, according to Indian Express, of the same day), comprising of demographic details of each of the persons vaccinated, captured from Aadhar database.

Here, in this paper, we shall review some notable methodologies and approaches to measure human/social development, as available from literature, and then suggest and develop one model to measure the regional diversity, for different districts of Odisha (in India), using relevant secondary data.

2 Review of Literature

One of the most popular methodologies to measure human development is HDI (Human Development Index). This HDI is calculated by taking several socio-economic factors, and the formula is differently computed by different scholars.

The Human Development Report (1990) emphasised three measures of human development, which are mentioned below, as summarised by Mahabub-ul-Haq:

(a) *Human Development Index (HDI):* giving importance to 'healthy life', literacy, living standard and GDP.
(b) *Gender (related) development index (GDI):* giving importance to female health and life-expectancy, female—literacy and their income.
(c) *Human Poverty Index (HPI):* consisting of factors like 'vulnerability to death', 'standard of living', 'percentage of malnourished children under five'.

For calculation of HDI, in his study, Tripathy [1] considered the indicators, like: Percentage of young school students (within the age group of 6–14 years), Percentage of male literacy, Percentage of female literacy, Number of hospital beds, for analysing district-wise human development in Odisha. Tripathy [1] found that human development is significantly affected by Life Expectancy Index (LEI), Education Index (EI), and Gross Domestic Product Index (GDPI) but out of them the high human development districts of Orissa are affected more by the GDPI. In the medium human development districts, EI is more weighted than LEI and GDPI. In case of the low human development districts of the state, EI has also more importance. Most indices have a negative relation with the HDI. There is huge number of errors in low human development districts as compared to high and medium human development districts.

The important factor is the health index, which influences human development in Orissa more as compared to other two factors like income and EI.

Foster et al. [2] in their study have taken State rankings, Changes in country rankings and Losses due to inequality as constructs and examined a new approach for integrating 'the distributive dimension into the HDI'. A new class of human development indices developed by them, in which, the traditional index and a family of indices both are included, which are sensitive to the distribution of human development. The general mean has been used by this class of indices to summarise achievements within each dimension of development, as well as to aggregate across dimensions.

Sharma [3] in his study examined the usefulness of HDI as a sensible measure of human development.

Acharya and Wall [4] have also contributed to the construction of HDI taking many unique indicators like environmental quality and human rights.

3 Mahalanobis Distance Statistics

Mahalanobis distance statistics was given by P.C. Mahalanobis in the year 1936, which was applied to measure the social development (distance) of different tribes (as compared to mainstream humanity). It is a distance between a point/vector (P) and a distribution (D). It is a multidimensional generalisation of the idea which measures the distance of a point from the mean of distribution by how many standard deviations. It has been applied for multivariate outlier detection, classifying the highly imbalanced (social) data sets and for subsequent classification.

Earlier, the distance between two points was measured by Euclidean distance. Mahalanobis distance statistics (or, D-square statistics) has a different approach, which can be illustrated by the following example:

Suppose there are two points in a 2D surface. It means it has two numerical columns, which are P and Q, in the given datasets. Therefore the distance between the two points, i.e. (P1, Q1) and (P2, Q2), can be stated as

$$\mathrm{d(P, Q)} = \sqrt{(P1 - Q1)^2 + (P2 - Q2)^2}$$

This measure can be extended as per the number of dimensions, like

$$\mathrm{d(P, Q)} = \sqrt{(P1 - Q1)^2 + (P2 - Q2)^2 + (P3 - Q3)^2 + \cdots + (Pn - Qn)^2}$$

It has two assumptions. First, the dimensions should be equally weighted and second, these should be independent of each other. It only measures the distance between two points. It doesn't consider the variation amongst the rest of the points.

The formula for calculating Mahalanobis distance is

$$\mathrm{D}^2 = (x - m)^T \cdot C^{-1} \cdot (x - m)$$

where,

D^2 Square of the Mahalanobis distance.
x Vector of observations (row in a dataset).
m Vector of mean values of independent variables (mean of each column).
C^{-1} Inverse covariance matrix of independent variables (mean of each column).

In a normal distribution, the probability of distribution is concave exactly at that region, where the Mahalanobis distance is less than 1.

Nowadays, many software and statistical packages, like 'R', Python, Julia, etc., are used to compute Mahalanobis distance, conveniently, using large social data.

4 Role of Social Statistical Information and Development Models

Development models in social science research have seen many empirical studies, investigating diversity amongst demographic divisions of populations. These studies majorly used social development indicators across multiple developmental periods and within community, caste, religion, genetic-lineage, and examined relationships of related constructs and behavioural outcomes. Tests of social development model hypotheses have demonstrated significant utility in understanding the causal factors and relates to socio-psychological behaviours. Development models, aimed at measuring the degree of diversity (or distance from the bench-marked standard), use different types of information (secondary data), called 'development indicators' (or constructs).

5 Sources of Social Statistics

Giving priority to social issues and human problems, the main data sources for this type of study are usually secondary databases published by census organisation and other agencies collecting various vital statistics. For the construction of indices like HDI, poverty index, social acceptance index or happiness index, we need to have relevant data on other socio-anthropological indicators.

Such vital statistics, including housing, agriculture, health, education, etc. can be obtained, in India, from sources like local bodies, hospitals, dispensaries and local self govt., like:

https://www.censusindia.gov.in/

http://wwwmospi.nic.in/

https://www.mygov.in/

https://www.lgdirectory.gov.in/

Administrative units, like revenue department (for land holding), judiciary (for legal support system), disaster management and mitigation units (for distress rehabilitation data), crime-control-research units (for crime against the oppressed and marginalised or minority class), etc., (like: https://www.ncrb.gov.in/en/crime-in-india-table-addtional-table-and-chapter-contents) are also very useful for getting relevant secondary data.

6 Social Support System and Regional Safety Status: A Case Study of Odisha Districts

We can take a very large number of units for index-based comparisons, like all districts or blocks of India, or all (96,37,820) households of Odisha, for which data are readily available. But for space constraints, and as an illustration of index construction exercise, we shall take 'safety and security of citizens', only for the districts of Odisha.

For ranking districts of Odisha, on the basis of 'social safety and security', secondary data on the rate of latest crimes against children and ladies were taken and compared with the population density of each district. This latest crime Report-2020 (released in September 2021, by NCRB, at: https://www.ncrb.gov.in) was considered.

Using data from Table 1, it was first studied to find if any association existed between the occurrence of crime and sex ratio, and with density. As expected, it was found (from Table 2) that there existed a positive correlation between population density and crime incidences (higher in more populous districts). It is interesting to note here that there is a negative correlation between the sex ratio and rate of crime. That is, the higher the sex ratio, less the crime (Table 2).

Table 1 Crime statistics of districts of Odisha

1	2	3	4	5	6	7
Crime against SC-ST	Crime against women	Crime against children	Districts of Odisha	Sex ratio	Density	Literacy
130	1246	228	Angul	943	200	77.53
131	1613	411	Balasore	957	610	79.79
98	497	106	Baragarh	977	254	74.62
88	884	336	Bhadrak	981	601	82.78
168	973	185	Bolangir	987	251	64.72
35	211	46	Boudh	991	142	71.61
144	656	184	Cuttack	940	667	85.5

(continued)

Table 1 (continued)

1	2	3	4	5	6	7
34	200	37	Deogarh	975	106	72.57
98	1091	217	Dhenkanal	947	268	78.76
13	268	93	Gajapati	1043	134	53.49
93	1002	132	Ganjam	983	430	71.09
107	1056	212	Jagatsinghpur	968	682	86.59
198	1411	341	Jajpur	973	630	80.13
66	436	118	Jharsuguda	953	274	78.86
80	842	187	Kalahandi	1003	199	59.22
26	352	115	Kandhamal	1037	91	64.13
112	1413	217	Kendrapara	1007	545	85.15
77	765	191	Keonjhar	988	217	68.24
38	554	126	Khurda	929	800	86.88
33	647	170	Koraput	1032	157	49.21
18	446	148	Malkangiri	1020	106	48.54
144	1490	448	Mayurbhanj	1006	242	63.17
46	508	156	Nayagarh	915	248	80.42
60	673	207	Nabarangpur	1019	231	46.43
60	240	82	Nuapada	1021	158	57.35
147	1544	274	Puri	963	488	84.67
42	547	185	Rayagada	1051	137	49.76
59	781		Sambalpur	976	157	76.22
65	173	170	Sonepur	960	261	74.42
54	465	38	Sundargarh	973	216	73.34

Note Estimated data for 2020, as available from the following sources
Source for column (1, 2, 3): https://www.ncrb.gov.in/crime-in-india-table-contents
Source for column (5, 6, 7): https://www.census2011.co.in/census/state/districtlist/orissa.html

Table 2 Correlation table

Correlation	
r(CAW-literacy)	0.37
r(CAW-Density)	0.513
r(CAW-Sex Ratio)	−0.21
r(CSCST-Literacy)	0.552
r(SCST-Density)	0.63
r(SCST-Sex Ratio)	−0.42

Note CAW = Crime against women, CSCST = Crime against SC-ST persons;
Correlation values computed by author using data from Table 1

Table 3 (a): Worst status states; (b): Better performing districts; (c): Average

(a) *Index > 1*	Worst status
4.454	Mayurbhanj
3.816	Angul
2.503	Sambalpur
2.049	Malkangiri
1.789	Rayagada
1.649	Koraput
1.617	Kalahandi
1.578	Kandhamal
1.327	Dhenkanal
1.059	Bolangir
(b) *Index < 1*	*Best status*
−2.746	Sonepur
−2.704	Khurda
−2.294	Cuttack
−1.744	Boudh
−1.524	Bhadrak
−1.516	Nuapada
−1.318	Deogarh
−1.179	Baragarh
(c) *1 < Index < 1*	
Ganjam	
Nayagarh	
Gajapati	
Jajpur	
Sundargarh	
Kendrapara	
Balasore	
Puri	
Nabarangpur	
Keonjhar	

7 Development of Social Security Status of Districts

Using crime data for each district, and other population statistics, an relative index is presented here, by adding all crimes and then dividing by population density, by the following formula:

SIID (Social Insecurity Index of Districts) = [(CAW + CAC + CA SCST)/ PD] –(μ),

where CAW = Crime against women, CAC = Crime against Children, CASCST = Crime against SC-ST, and PD = Population density, μ = State average ratio = Total Crime/State PD.

The following table, presented in three blocks, gives 'best (with negative score)', 'worst (with positive scores, arising because of the higher number of crimes in the denominator)' and average (or status-quo) status of districts of Odisha. The same methodology can be applied to find the national average and then all districts of India.

Above small illustration, just for Odisha districts, used secondary data of 77 × 35 matrix size. We can suggest a similar exercise for other social development (secondary) data, with a very large set of indicators. Following illustrative conceptual model is suggested for possible construction of SWI (Social Wellbeing Index). Odisha State Government administration should look into the state of affairs at distorts like, Mayurbhanj, Angul, Sambalpur, and Malkangiri. Things are looking really good for the districts of Sonpur, Khurdha, Cuttack, Boudh, and Bhadrak.

8 Conceptual Model for Constructing Family Wellbeing Model

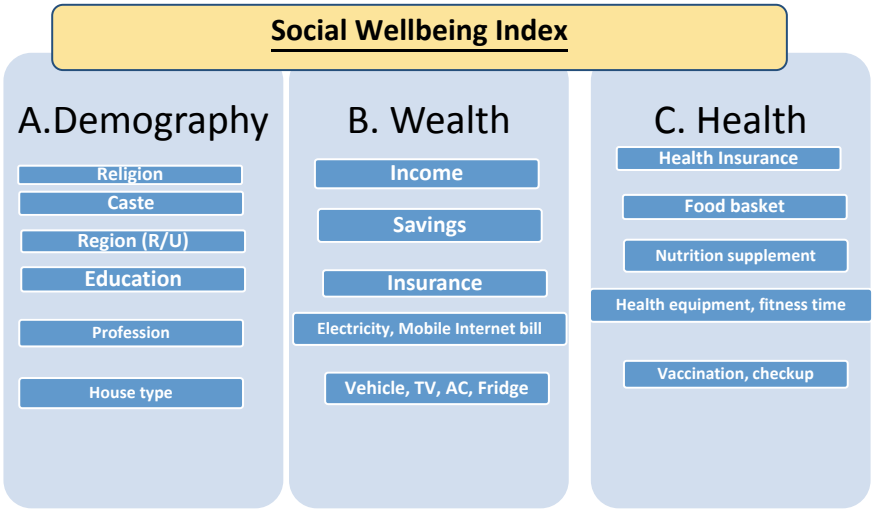
To understand and compare the relative status of family wellbeing, at the household level, we may consider following socio-economic and demographic factors and can get data from various sources, as mentioned above. This can further be expanded and have more operational variables/indicators.

Input Variables/Indicators

Religion|Caste|Region (rural/Urban)|Income|Education|Profession|House-type|Vehicles|Insurance|Food basket budget| Savings|Water source|Health/ vaccination|Electricity Bill||LPG Bill||News magazine Bill||Mobile-internet bill| TV, Fridge, AC availability|.

Output/Criterion Variable

Social Wellbeing Index (and perception scores).



[Author's own suggestion]

9 Conclusion

Studying and constructing social development indices are of great social importance, having practical use value for policymakers, planners and administrators. It can indirectly help measuring the impacts of intervention programmes. As found from this illustrative study, we can observe better gender ratio can have a positive role in checking crimes and creating a better social security system. In a similar way, as a possible future research suggestion, from the above conceptual model on the construction of 'social wellbeing index', we can possibly find the relative importance of smaller, yet newer indicators, like vaccination and health insurance plans. We suggest that, more and more such measures should be initiated, we relatively newer social indicators, using various scales.

There are many other virgin areas, where indices similar to HDI could be applied.

Acknowledgements The author acknowledges the help and contribution of his research scholar, Mr. Smruti Malhar Mohapatro, for the collection of data and other secondary information, from the above cited sources, identified by the author.

References

1. Tripathy, U.: Estimation of human development index in Orissa: district-wise analysis. IUP J. Manag. Econ. **8**(4) (2010)
2. Foster, J.E., Lopez-Calva, L.F., Szekely, M.: Measuring the distribution of human development: methodology and an application to Mexico. J. Hum. Dev. **6**(1), 5–25 (2005)
3. Sharma, S.D.: Making the human development index (HDI) gender-sensitive. Gend. Dev. **5**(1), 60–61 (1997)
4. Acharya, A., Wall, H.J.: An evaluation of the United Nations' human development index. J. Econ. Soc. Meas. **20**(1), 51–65 (1994)
5. Tripathy, U.: Estimation of human development index in Orissa: district-wise analysis. IUP J. Manag. Econ. **VIII**(4) (2010)

Web-Source

6. https://www.odisha.data.gov.in/catalog/households-marital-status-sex-and-age-head-household-india-and-states#web_catalog_tabs_block_10

Deep Learning Trends and Inspired Systems in Natural Language Processing

Aseer Ahmad Ansari, Siddharth S. Rautaray, and Manjusha Pandey

Abstract Text data is rapidly becoming a commonplace entity. Social media forms a nebula of such data that is easily accessible to common people and researchers. For corporate and other businesses, its surveys and company mails that provide them with the data. So it is not a surprise for the field of Natural Language Processing to witness a consistent rise in research and insights and more specifically in Natural Language Understanding (NLU) by precepting syntax, structure and sentences together. Recent advancements in representation learning methodologies have also unlocked understanding of text greatly by exploiting a common and overlooked point of view, i.e., attention. This research chapter will explore the recent trends in deep learning and the inspired systems in this area by charting the insights and inspiration that have come to build our current state-of-the-art models and architecture. This work also presents a summary of similarities and contrasts of various such models to conclude on this evolution of deep learning in NLP.

Keywords Deep learning · Attention · Embeddings · Transformer · RNN

1 Introduction

Natural Language Processing (NLP) or Computational Linguistics is the processing of human language through computational techniques. It enables computers to perform a variety of tasks at many levels. Text parsing and understanding, parts of speech tagging, sentiment analysis, machine translation, and many more. Recently there have been many advances in this field extending solutions to areas like biomedical research and clinical and genomic diagnostics. The methods in deep learning

A. A. Ansari
Deloitte India LLP, Chennai, India

S. S. Rautaray · M. Pandey (✉)
School of Computer Engineering, KIIT Deemed to Be University, Bhubaneswar, Odisha, India
e-mail: manjushafcs@kiit.ac.in

S. S. Rautaray
e-mail: siddharthfcs@kiit.ac.in

S. S. Rautaray et al. (eds.), *Data Science in Societal Applications*, Studies in Big Data 114, https://doi.org/10.1007/978-981-19-5154-1_5

which aim at learning representations in data through the processing of multiple layers have been at the core of producing the current state-of-the-art results in this field and others. This has been in contrast to the machine learning models like SVM or decision trees that have been around for a long time and are majorly trained on high dimensionality based sparse features and most often such machine learning developed systems were dependent on hand-crafted features which resulted in them being time consuming and mostly inferior.

Presently, research in NLP has evolved up to the era of google search being able to process and return accurate results after processing millions of documents and webpages in less than a second. This has been the result of reliance on new deep learning methods that are reaching newer and higher benchmarks more rapidly than ever. These methods like dense vector representation learning, successful word embeddings, and others are the reasons for said superiority. Collobert et al. in [1] have demonstrated results from a simple deep learning framework easily outperforming many classical approaches in a variety of downstream tasks like POS tagging, named entity recognition, and semantic role labeling. And so many complex deep learning models have since then solved the more complex NLP tasks ranging from Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and recursive neural networks.

In this paper, the work is segregated into sections beginning with an introduction to the concept of distributed representation and its aid in traditional machine learning models and other sophisticated deep learning models.

2 Word Embeddings

Word embeddings are developed on the hypothesis of word similarity and such words occurring in a similar context. Hence based upon this, embeddings of words capture those contexts from their neighbors. This ability is developed into an advantage for the embeddings since they can capture word similarity. Distance metrics like cosine similarity or euclidean similarity are used to capture that as shown in Fig. 1. In traditional machine learning models, these have often replaced the more classical TF-IDF (term frequency-inverse document frequency) up to a significant degree of advantage in many NLP tasks. Deep learning models on the other hand make use of these embeddings as mostly their first level of distributional embeddings vector.

So far, these word embeddings have typically been trained on a huge unlabeled corpus and optimizing an auxiliary objective like predicting the next word in context. As a result, word embeddings have been successful in efficiently storing word similarities, and then optimizing on them as per the nature of the context. Another advantage so far is owed to the small dimensionality of these embeddings enabling faster and more efficient training time in various NLP core tasks. This is the result of shallow networks behind these embeddings and being responsible for state-of-the-art results in a variety of tasks hence no there hasn't been a requirement for deep networks to train any better vectors.

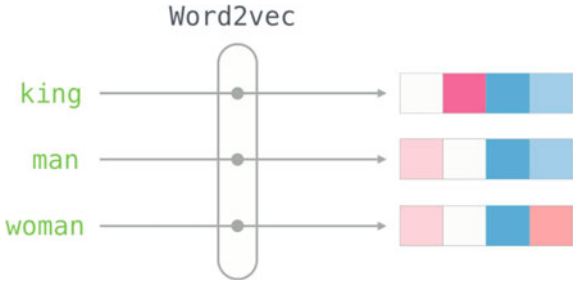


Fig. 1 Word2Vec embedding representation. *Source* Jay alamar

While many varieties of representations have been developed so far, the foundation for these distributional semantics was laid by Elman et al. [2] and developments in the 1990s. Many later developments have been the adaptations of these. For instance, [3] used stacking of denoising autoencoders along with these embeddings in sentiment classification for adapting to topics and domains. Another work by Hermann et al. [4] learned sentence composition by using a combination of categorical autoencoders. These instances and many more extended works in areas of topical modeling like latent Dirichlet allocation [5] and language models [6] are evidence of the wide range of acceptance and effectiveness of such embeddings.

Beginning with Bengio et al. work [7] in 2003, a neural language model was proposed that learned word representations and allowed for generalization as a result of which new and unseen sentences could result in higher confidence if a similar word sequence were already seen. The utility of pre-trained embeddings was first illustrated by Collobert et al. [6] who proposed a novel network architecture that forms the foundation of many present-day works. But it can be argued that the popularity of such high-quality embeddings was due to Mikolov et al. [8] who introduced the Continuous Bag of words and the skip gram model. An unexpected result of these models was their compositionality, i.e., adding 2 vector words could result in a vector that is semantically composite. For example, adding vectors of the two words "man"+"royal" would result in a vector semantically similar to "king". This was later explained in the works of Gittens et al. [9] as a result of uniformly distributed words in the vector space.

Glove [10] and Word2vec [8] are other famous word embedding models. Also this form of token embedding which began with words was followed by character level, byte level, and sub-word tokenization. The results from these are now incorporated into more complex deep learning models that will be discussed in depth in Sect. 4.

2.1 Embedding Models

The revolution in word embeddings was unarguably marked by Mikolov et al. [8, 11] who proposed the skip gram and the CBOW model. Both the model's functions complement one another. While the CBOW model target is to calculate the

conditional probability of a word given its neighboring window is the size of k, the skip gram model predicts the surrounding context words given a central target word. These context words are symmetrically located in the neighboring window of a predetermined size. The size of this window can be different by the task but the embedding dimensions are usually chosen in an unsupervised setting. The accuracy of the prediction in such a task could determine the optimal embedding dimension. For instance, increasing the dimensions of the embeddings consequently increases the accuracy up until the function converges and this embedding can then be considered as the smallest one without compromising on the accuracy. The CBOW model unlike that the standard bag of words model uses distributed representation of context with the following training complexity

$$Q = N \times D + D \times \log2(V)$$

Skip Gram architecture is proportional to

$$Q = C \times (D + D \times \log2(V))$$

And since the complexity of this model increases with a higher context range, a smaller window size like 5 was selected randomly with the assumption that the more distant words will atypically be related to the target word. Table 1 shows the results from a single CPU trained model on publicly available word vectors.

Since the word embeddings have a greater dependency on the application context, the following 2 solutions were proposed later. Using task specific embeddings [12] which retain the current task space could keep the model efficient but since training the embeddings from scratch is a time-consuming process it wasn't very practical. The quality of these vector representations and their training speed was then later improved by better sub-sampling methods and by the introduction of Negative Sampling (NEG) [8] as an alternative to hierarchical softmax. This was inspired by

Table 1 Full vocabularies model comparison on publicly available Semantic-syntactic Word Relationship test set

Model	Vector dimensionality	Accuracy (%)		
		Semantic	Syntactic	Total
Turian NNLM	50	1.4	2.6	2.1
Turian NNLM	200	1.4	2.2	1.8
Mikolov RNNLM	80	4.9	18.4	12.7
Mikolov RNNLM	640	8.6	36.5	24.6
Mnih NNLM	50	1.8	9.1	5.8
Mnih NNLM	100	3.3	13.2	8.8
CBOW	300	15.5	53.1	36.1
Skip gram	300	**50.0**	**55.9**	**53.5**

Table 2 Skip gram models result with and without sub-sampling on phrase analogy dataset containing 1 B words

Method	Dimensionality	10^{-5} Subsampling (%)	No subsampling
NEG-5	300	27	24
NEG-15	300	42	27
HS-Huffman	300	**47**	19

the application of Noise Contrastive Estimation (NCE). The efficiency increases as a result and as an extension of it, Liu et al. 2015 illustrate the impact of adding prior information into the skip gram model. The Topical Word Embedding Model (TWE) which combined text words and the context topic derived from LDA to compute the embeddings, was a result of treating the topical information as important prior knowledge (Table 2).

While the word embeddings model improvement was still going on, a common problem was observed in a corpus of huge vocabularies which was the unknown word issue. They were also referred to as Out-Of-Vocabulary (OOV) words. Another reason prominent was for tasks like POS tagging and NER where shape information and intra-word morphology are useful. Thus character embeddings naturally solves for OOV words since each word is eventually a combination of the character vocabulary of the language. These embeddings have also attracted many research prospects too. For instance, Kim et al. [13] used a CNN +highway character network with outputs into an LSTM RNN-LM and it achieved the same state-of-the-art results on Penn treebank with 60% fewer parameters suggesting the sufficiency of character embeddings for many language models. Other works like [14–16] used the same for improving performance or decreasing parameters in previously well-performing word embedding dependent models. Moreover, languages like Chinese are composed of individual characters and their semantic meaning is formed from their composition, hence character embeddings are automatically the better choice to begin with. This trend has been extended to other such similar languages like French, German, Spanish, Arabic, Czech, and Russian for employing deep learning models for NLP tasks.

Many other approaches have also been researched so far to address the drawbacks of word embeddings. Initialization of unknown words from the concatenation of their context words and then optimizing their embeddings by a greater learning rate was proposed by Herbelot et al. [17] but it is not conclusively proven to be as effective yet for major NLP tasks. Another approach by Pinter et al. [18] recreated pre-trained vector embeddings by training a character-based model. This was interesting in the sense of learning compositions of word vectors from their character vectors. But even after this line of research Lucy et al. [19] found many other limitations in terms of the word vectors actually being able to capture the required facets of contextual meaning and suggested that distributional vectors alone cannot capture these concepts behind the words. Furthermore, these vectors carried another inherent disadvantage, which was their inability to capture the word order.

Positional Embeddings were later introduced with the idea of encoding a finite-dimensional representation of the location of the words in sequence. They went from being an absolute position vector for the sequence to being represented by a positional matrix each row of which would be an interpolated position of the discrete values associated with the item of that row index. In [20] the authors showed that those positional embeddings capture only the position of individual words [21] and not the relationship like precedence or adjacency between separate word positions. They proposed a continuous word function that would then replace the independent vectors and would smoothly capture the representation while shifting as the position increases. Also, since absolute position cannot be accessed in the middle layers, relative positioning can be pre-calculated and fed as an additional embedding vector to the NN. As a result, word representations in different positions can correlate with other words through this function. This was also proved by Gehring et al. [21] by modeling distance between elements in the sequence as well as an explicit use of relative position encoding to incorporate word distance relationships.

The more recent frameworks that have replaced recurrent and convolutional functions much like ConvSeq [21] and the transformer model [22] have introduced in their architecture an additional positional embedding in the feature level itself. The attention mechanism that is integral in this replacement is discussed further in depth in Sect. 3 and the models that result in the present state-of-the-art results are also collectively compared and summarized in Sect. 4.

3 Neural Mechanisms and Attention

Convolutional, recurrent, gated recurrent, and long short-term memory neural networks have conclusively been proven as state-of-the-art architectures in modeling sequence and problems of transduction and others like language modeling or machine translation, etc.

3.1 Recurrent Neural Networks

RNN specifically processes sequential information. Their recurrent nature indicates iteration over the same task on each sequence for them to have outputs dependent on their preceding inputs and the results. This allows the recurrent units to in a manner store the memory from previous computations and consequently use this information in further processing. This architecture has suited all NLP tasks such as speech recognition, image captioning, language models, and machine translations.

Since RNN performs sequential processing, it can capture the inherent sequential nature of a language as shown in Fig. 2. The units in a sequence could be words, phrases, or characters. Because of this inherent sequential dependency parsing nature, they are naturally more suited and better considered than CNNs.

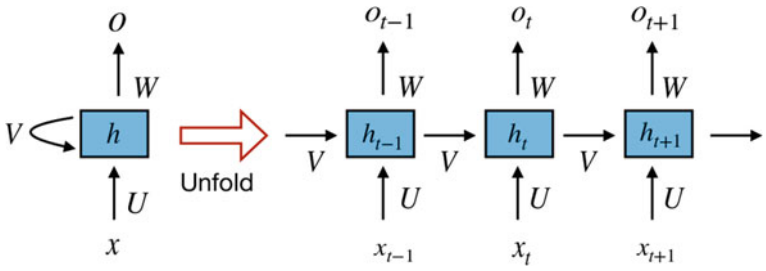


Fig. 2 Simple RNN. *Source* leCunn

Another advantage is their ability to model input of variable length, since many tasks in NLP require arbitrary length sentences which are given to their flexible steps for computing and capability to capture unbounded context. Their distribution joint processing aids applications like POS taggings [23], while their summarization [24] capability helps in machine translations. In deeper vanilla RNN models, an inherent disadvantage called vanishing gradient disables the tuning and learning of previous layers parameters. LSTM, gated recurrent unit, and residual networks solved this infamous problem.

LSTM makes use of three gates to calculate the hidden state and this allows the model to back-propagate error through almost infinite time steps. Gated recurrent units that were empirically designed are a similar variant but are less complex computationally and hence are sometimes viewed with an advantage over LSTM.

While both the mechanisms work amazingly well, researchers have not been able to conclusively call out the better gating mechanism. Chung et al. [16] performed an extensive comparison but could only conclusively show the superiority of the gating mechanism over the vanilla RNN and not classify the better of the two gates.

3.2 Convolutional Neural Networks

In accordance with the NLP domain being dominated by sequences, it was never assumed to be a task fit for convolutional neural networks. However, several researches indicated some conflicting evidence over the superiority of CNN over RNN. The results spoke for themselves as in tasks suited naturally well for an RNN like language modeling, CNN caught on with a better performance. In principle both the models deal differently with sequences while modeling them. In addition to dealing with unbounded context, RNN creates compositions of variable length sentences. CNN on the other hand models the sequence in order to derive the most important parts from the sequence which in some cases is a disadvantage as while the practice is sufficient to capture the important feature in a restricted frame for some classification tasks, it ignores the longer dependencies in the sequence. A study by Yin et al. [25] extracted interesting insights while comparing the performance of

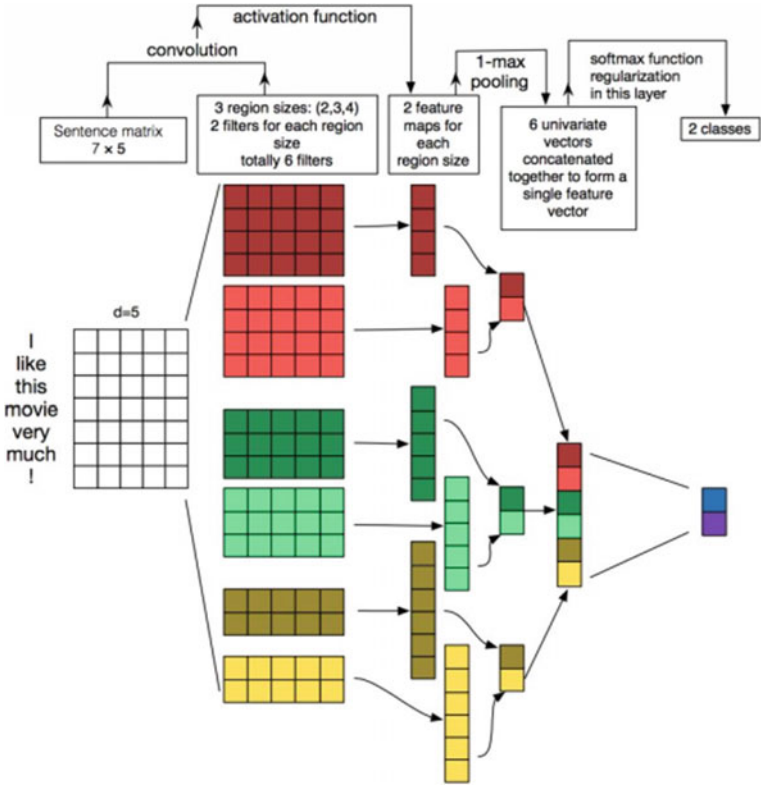


Fig. 3 Text modeling by a CNN

both the models on various different tasks such as QA, POS tagging, and Sentiment classification but couldn't indicate a clear winner.

CNN ability has been demonstrated to extract important n-gram features with models tracing back to [1]. After this various applications [6, 26, 27] were researched by post which actual CNN proliferation was found in NLP literature. Figure 3 shows a simple CNN modeling text for classification.

3.3 Attention Mechanism

The attention mechanism is among the most valued breakthroughs in the field of deep learning in the last decade. While it has given rise to even more breakthroughs in all fields, the major ones have been in NLP which includes architectures like transformers. The first emergence of the attention mechanism was the result of improvements in the encoder-decoder-based machine translations.

Neural machine translations replaced the phrase-based translations [28] as proposed by Kalchbrenner et al. [29], Sutskever et al. [30], and Cho et al. [31]. The earlier approach worked to fine tune the smaller components separately, while the neural approach worked to collectively build and train a single neural network. Most of these neural networks were based on the encoder-decoder architecture which worked in two parts. First the encoder would encode the input information and in a sense summarize it into a fixed length vector and in the second step the decoder would use this output to make the translations and the whole system's objective would be to maximize the joint probability. Consequently the quality of the result of translations produced by the decoder part would be only as good as the encodings. A problem with this architecture was its inability to capture and encode long-term dependencies in the longer length sentences, since the objective was always to encode the information into a single fixed length vector. [31] also showed that with the increase in sentence length the performance of this architecture decreases.

Bahdanau et al. [32] addressed this issue by introducing what is now called the "attention mechanism." It was proposed as an extension to the encode-decoder architecture itself to learn the alignment and jointly translate it into the model itself. In principle, once the model would generate a translation of a word, a soft search would look for items and identify them in the sequences that would be the most important ones or essentially where the information is concentrated. Figure 4 shows a general illustration of the proposed mechanism.

In this new proposal, they define each conditional probability earlier defined as

$$p(y) = \prod_{t=1}^{T} p(y_t | \{y_1, \ldots, y_{t-1}\}, c),$$

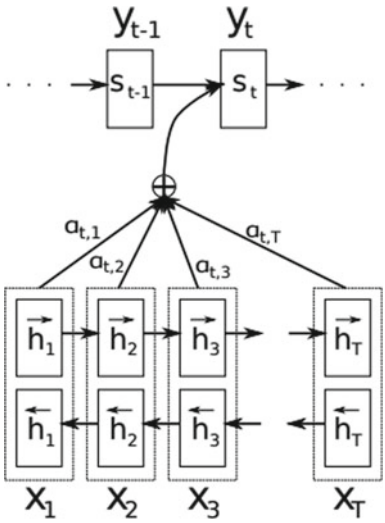


Fig. 4 A graphical illustration of the proposed attention model

to,

$$g(y_i - 1, s_i, c_i)$$

Here, si is the hidden state of the RNN for the time stamp i defined as

$$s_i = f(s_i - 1, y_i - 1, c_i)$$

As shown, the probabilities are now conditioned on ci which is a distinct context vector for each word y_i and these context vectors now depend on the sequence of annotations $(h_1, \cdots, h_{Tx})$ as shown in the figure which now contains the contextual information of the entire sequence as well as information on neighboring context surrounding the ith item. The model would then predict the target based on these context vectors. This model also helps with another problem which is the fixed length vector, since now unlike the basic encoder-decoder it doesn't attempt to encode the entire sentence and instead encodes the input into a sequence of vectors, while adapting a subset of these vectors will then be chosen by the decoder. This clearly solves the problem of prematurely compressing information of the entire input sentence irrespective of the length and provides the model with the freedom to optimize the decoding process from the important input information. The research showed that the results were mostly apparent as the sentence length increased but could still be observed with any variable length vectors. Figure 5 shows the RNN model's BLEU scores with both the general and the proposed search.

Also, on an English-to-French translation task this model easily achieved a close and comparable performance to the earlier phrase-based translation system. This architecture clearly formed a promising step forward. Table 3 shows RNNsearch clearly outperforming conventional RNN as measured through their BLEU scores. The proposed model performance compares to a phrase-based conventional system called Moses. This is to show a significant performance increase given that the conventional system uses a separate monolingual corpus besides the corpora used to train their own model.

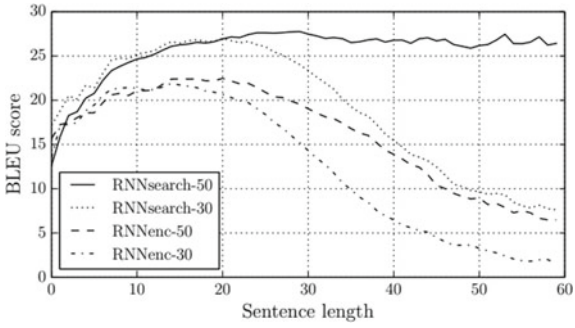


Fig. 5 BLEU scores of the translations form proposed models with different sentence lengths

Table 3 BLEU scores model comparison on the translations test set. RNN search-50* was trained until the performance peaked

Model	No (UNK)	All
RNN encdec-30	24.19	13.93
RNN search-30	31.44	21.5
RNN encdec-50	26.71	17.82
RNN search-50	34.16	26.75
RNN search-50*	**36.15**	28.45
Moses	35.63	**33.3**

The proposed approach also allows us to visualize and inspect the generated translations and the source words as shown in Fig. 6.

The next challenges that remain in this model were to better handle the unknown or rare words and also tackle another inherent disadvantage of the sequential nature of the model which is precluding parallelization. At longer sequence length the memory constraints become more concerning and limit the batches across training. Although there has been researching to improve this computational efficiency such as through conditional computation [33] or factorization [34], inside the model sequential issue still persists. Vaswani et al. [35] demonstrate and emphasize the goal of decreasing the sequential computation. This is the transformer which is briefed in the next Section. The attention mechanism used in it is referred to as "self-attention." This mechanism relates to different positions in its own sequence to then eventually compute a representation of its sequence. Their proposal includes what they call a Scaled Dot product attention which queries, keys and values. The attention function

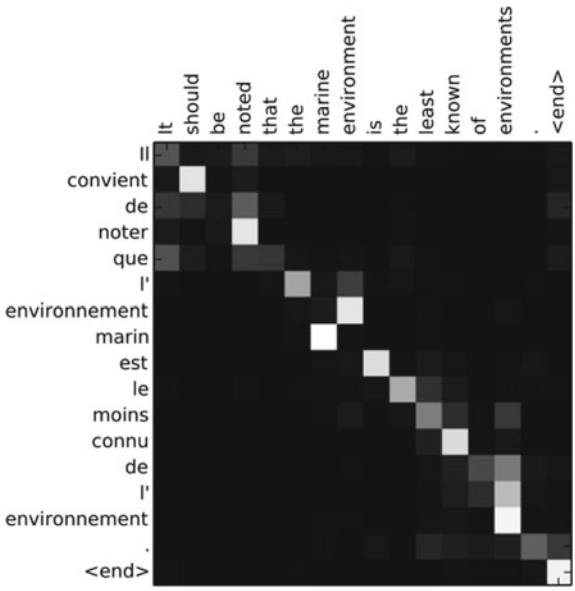


Fig. 6 x and y axis show source (English) and target (French) and each pixel shows the weight of the annotation

Table 4 Different layer type comparisons. n is sequence length, k is convolution kernel size, r is neighborhood size in restricted self-attention, d is embedding dimension

Layer type	Sequential operations	Maximum path length	Complexity per layer
Self-attention (restricted)	$O(1)$	$O(n/r)$	$O(r.\ n.\ d)$
Self-attention	$O(1)$	$O(1)$	$O(n^2.\ d)$
Recurrent	$O(n)$	O(n)	$O(n.\ d^2)$
Convolutional	$O(1)$	$O(log_k(n))$	$O(k.\ n.\ d^2)$

is then calculated as a dot product of the queries with all the keys as shown below:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The dot product is used in their proposal instead of additive attention since the latter is much faster and space efficient as well. The other difference from the standard attention is scaling the dot product by a factor $1/\sqrt{\mathrm{d_k}}$ since generally additive attention performs better for larger values $\mathrm{d_k}$, since the softmax outputs extremely smaller gradients when dot product magnitude is large. Hence the scaling helps in counteracting this effect.

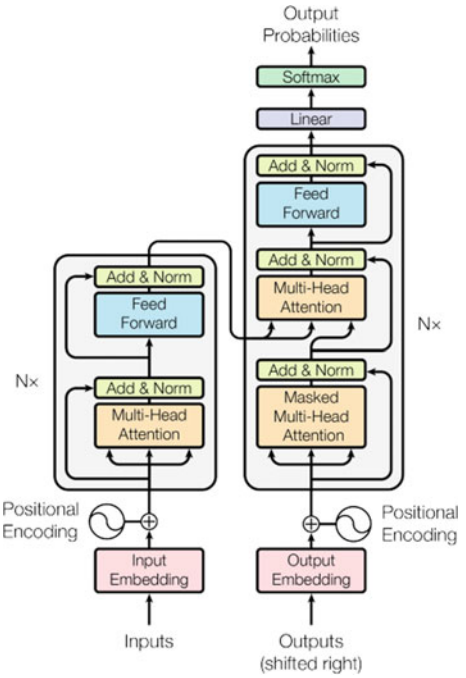
A complexity comparison for self-attention is shown in Table 4.

4 Transformers

At this point, the attention mechanism has already become an integral part of sequence modeling tasks and for some computer vision applications too. Attention mechanisms were never alone controlled or computed but rather were combined with operations like recurrence or convolutions. The work by Vaswani et al. [35] proposes the Transformer which does away with any form of recurrence and uses the attention mechanism entirely to compute global dependencies. This kind of architecture even allows for significant parallelization and produces state-of-the-art results within less training time. In contrast to a standard encoder-decoder architecture or the one proposed by Bahdanau et al. [32], the number of operations would increase directly with the distance between positions and as a result learning these dependencies was already difficult between difficult points. The proposed transformer model reduces these operations to be constant and not dependent on the item distances. This approach causes a reduction in effective resolution but this problem is dealt with by a new approach proposed in the same research called multi-head attention.

The overall transformer architecture is made up of stacked self-attention and fully connected encoders and decoders. Figure 7 shows this architecture.

Fig. 7 Transformer Model



Encoder: 6 identical layers are stacked and each layer is composed of point-wise fully connected network and multi-head self-attention. A residual connection goes around every 2 sublayers and is followed by normalization.

Decoder: They are again the same 6 layers stacked with sublayers including one more operation which is to perform multi-head attention operations for the outputs from the encoder stack. The self-attention layer is modified a little in the decoder but the other residual connections and normalizations are the same as in the encoder.

The self-attention in this transformer architecture addresses three problems collectively now (Table 5).

1. Reduce Computational Complexity per layer
2. Parallelization of computation or reduction of sequential computation.
3. Less computations by path length for lengthier dependencies.

The Transformer model paved the way for many attention-based models to come and which are not restricted to text but extend to efficiently handle large data inputs of audio, images, and videos.

Devlin et al. [36] introduced the bidirectionality in the transformers and proposed a model called bidirectional encoder representations of transformers (BERT). This model would pre-train representation containing both left and right context and thus storing in those representations a bidirectional context. Being a pre-trained model this would require just one more layer to tune the outputs for a variety of tasks. Unlike other models like ElMo which produce feature-based approaches to their pre-trained

Table 5 Performance of Transformer as measured by their BLEU scores on English-to-German and English-to-French newstest2014 in comparison to the previous state-of-the-art result models

Model	BLEU	
	En-Fr	En-De
ByteNet		23.75
Deep-Alt+PosUnk	39.2	
GNMT+RL	39.92	24.6
ConvS2S	40.46	25.16
MoE	40.56	26.03
Deep-Alt+PosUnk ensemble	40.4	
GNMT+RL ensemble	41.16	26.3
ConvS2S ensemble	**41.29**	26.36
Transformer (base)	38.1	27.3
Transformer (big)	**41.8**	**28.4**

model, BERT provides a fine tuning approach much like the OpenAI GPT [37]. Standard transformer model had one major limitation which is being unidirectional and as a result choice of architecture is limited for pre-training. BERT surprisingly introduces bidirectional using what they call a "Masked language Model" which is inspired from [38]. In this approach, they would randomly mask tokens in the corpus with the objective to predict the original ones thus enabling a vector representation that would combine left and right context. They demonstrate that the standard approach so far is nothing but a shallow concatenation of separately calculated bidirectional context. Their approach advances the state-of-the-art for a total of 11 downstream tasks. Another principle that is modeled into BERT's architecture is the ability to understand the relationship between 2 sentences and this is achieved through their objective of Next Sentence Prediction (NSP) and is an extension of the objective used in [39].

The results from BERT were amazing but the entire transformer routine is exhausting and gets costlier with longer sequences. So Kitaev et al. [40] introduce 2 techniques to improve the transformer's efficiency. Reversible residual layers and replacing dot production attention with Locality Sensitive Hashing (LSH). Their techniques solve the memory issues in the following manner:

1. A single copy of activation for the entire model can be stored by using reversible layers.
2. Memory is saved by splitting the activation and processing them in chunks.
3. Approximating the attention by using LSH and reducing the complexity from $O(L^2)$ to $O(Llog L)$.

They show negligible performance drops even after incorporating these changes (Table 6).

Thus reformed formed a transformer with efficient memory capabilities even on longer sequences. A similar approach was introduced by Lan et al. [41] to reduce the parameters of the transformers and allow better scaling compared to the original

Table 6 Shows the BLEU scores ewstest2014 for WMT English-German (EnDe)

Model	BLEU	SacreBLEU	
		Uncased	Cased
Transformer (base)	27.3		
Transformer (big)	28.4		
Reversible transformer (base, 100K steps)	27.6	27.4	26.9
Reversible transformer (base, 500K steps)	28.0	27.9	27.4
Reversible transformer (big, 300K steps)	29.1	28.9	28.4

BERT. This model is called ALBERT performer factorization of embeddings and secondly allows for cross-layer parameter sharing. Their configuration demonstrated BERT-like performance with 18× less parameters and 1.7× faster training time. They also propose to replace their standard NSP with what they called SOP or sentence order prediction. This essentially addresses the inefficiencies of NSP demonstrated in [42] and also takes care of the inter-sentence coherence. As a result of this introduction when the training time for the models is kept the same ALBERT outperforms other models on most tasks as shown in Table 7.

Many newer transformer-like models are being worked upon and more often than not the orthogonal line of research is to reduce computation and make the transformers more efficient. It is still hypothesized that there could be possibly more dimensions that are not captured yet which would allow more representational power and even better performance.

Table 7 Public data performance comparison of ALBERT and other transformer-based models

Models	MNLI	RTE	SST	CoLA	STS	WNLI
BERT-large	86.6	70.4	93.2	60.6	90.0	
XLNet-large	89.8	83.8	95.6	63.6	91.8	
RoBERTa-large	90.2	86.6	96.4	68.0	92.4	
ALBERT (1M)	90.4	88.1	96.8	68.7	92.7	
ALBERT (1.5M)	**90.8**	**89.2**	**96.9**	**71.4**	**93.0**	
ALICE	88.2	83.5	95.2	**69.2**	91.1	80.8
MT-DNN	87.2	86.3	96.5	68.4	91.1	89.0
XLNet	90.2	86.3	96.8	67.8	91.6	90.4
RoBERTa	90.8	88.2	96.7	67.8	92.2	89.0
Adv-RoBERTa	91.1	88.7	96.8	68.0	92.4	89.0
ALBERT	**91.3**	**89.2**	**97.1**	69.1	**92.5**	**91.8**

5 Other Works

NLP tasks such as language generation can be modeling into a problem involving an agent to perform discrete action and so Reinforcement learning can be casted to some NLP applications. Typically the objective of RNN or transformer models is to maximize the likelihood of each item in the sequence given its current and previous hidden state. Now at test time, the actual tokens are replaced by the token generated by the model and this is supervised by "Teacher forcing" at the time of training. This gap between training and inference is a discrepancy called "exposure bias". Another discrepancy observed is the gap between the training objective and the test metric.

Reinforcement learning shows the potential to address these problems up to an extent. Ranzato et al. [43] in an experiment to optimize the non-differentiable metric applied a reinforcement learning algorithm to train RNN models over a variety of language generation tasks. They viewed RNN as an agent interacting with the external factors like words and context representations while the model's parameters are assumed to be the agent's policy and bringing those policies into action would mean predicting the next word in the sentence. After successfully predicting the sentence until the end token the agent receives a reward. Li et al. [44] defined three different types of awards based on the flow of information, answering ease, and semantic coherence.

Deep Generative models have also been sought, given their recent success in being able to generate realistic images. Hence some effort is being applied to find its efficacy with text data. The objective would be to generate realistic sentences from a latent space and understandably discover rich information structure in the language.

Variational Autoencoders (VAE) [45] and Generative Adversarial Networks (GAN) [46] have also been reviewed recently to achieve the same objective. One problem that VAE could potentially solve is by imposing a prior distribution allowing the model to draw proper samples while the standard autoencoders do not impose any constraint on the latent space. As a result, the generated sentences are not realistic because mostly these sentences occupy a small region in the hidden spaces and none of them map to any realistic sequence. Bowman et al. [47] proposed a RNN-based VAE generative model that contained distributed representations of the entire sequence. Figure 8 shows their proposed VAE and sampling over the prior produce well-formed and diverse sentences.

GANs are an interesting combination of a generative and discriminating network. The generative network would decode the sequence representation into a data instance and the other discriminative networks distinguish between true and the synthesized data instances. But GAN does not work well in backpropagating the gradients through discrete variables. Even after the solution from [48, 49] the evaluation of deep GAN is a challenging process since for quality measurement similar data is required to compare with the unseen real data.

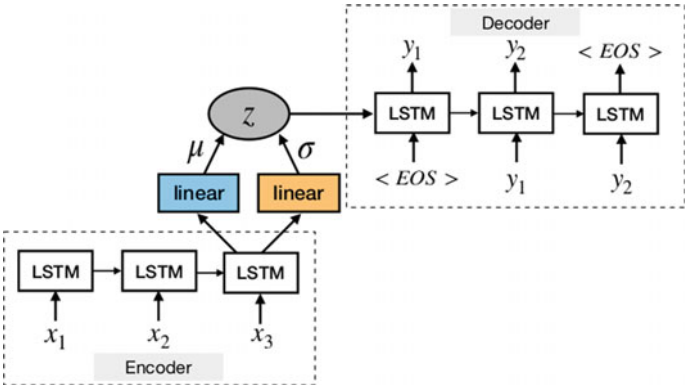


Fig. 8 VAE model. *Source* [45]

6 Future Scope and Conclusion

Deep Learning has provided us a way to use data and computation to exploit information and logic with much less effort. Newer models and concepts are continuously being researched and they are breaking benchmarks and setting the state of the art. Supervised learning branch of deep learning is undeniably the most popular but given the unavailability of labeled data in the real-world scenarios of text domain, the unsupervised and semi-supervised approaches are the best, and fortunately recent strategies of zero-shot learning and self-attention are the best employed. But these techniques are yet to be optimized and carved to their best potential. NLP domain research tangent is in the line of unlabeled data and there is much evidence to assume that the research is going to proceed in the same line for the time to come. We should expect to see better memory managing models particularly those whose internal memory is enriched with their external memory.

References

1. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**, 2493–2537 (2011)
2. Elman, J.L.: Distributed representations, simple recurrent networks, and grammatical structure. Mach. Learn. **7**(2–3), 195–225 (1991)
3. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: a deep learning approach. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 513–520
4. Hermann, K.M., Blunsom, P.: The role of syntax in vector space models of compositional semantics. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics (2013)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

6. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. ACM, pp. 160–167 (2008)
7. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
9. Gittens, A., Achlioptas, D., Mahoney, M.W.: Skip-gram-zipf+ uniform= vector additivity. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, 2017, pp. 69–76
10. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: EMNLP, vol. 14, 2014, pp. 1532–1543
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013). arXiv:1301.3781
12. Labutov, I., Lipson, H.: Re-embedding words. ACL (2), 489–493 (2013)
13. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. In: AAAI, 2016, pp. 2741–2749
14. Dos Santos, C.N., Gatti, M.: Deep convolutional neural networks for sentiment analysis of short texts. In: COLING, 2014, pp. 69–78
15. Dos Santos, C.N., Guimaraes, V.: Boosting named entity recognition with neural character embeddings (2015). arXiv:1505.05008
16. Santos, C.N., Zadrozny, B.: Learning character-level representations for part-of-speech tagging. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014, pp. 1818–1826
17. Herbelot, A., Baroni, M.: High-risk learning: acquiring new word vectors from tiny data (2017). arXiv:1707.06556
18. Pinter, Y., Guthrie, R., Eisenstein, J.: Mimicking word embeddings using subword rnns (2017). arXiv:1707.06961
19. Lucy, L., Gauthier, J.: Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning (2017). arXiv:1705.11168
20. Wang, B., Zhao, D., Lioma, C., Li, Q., Zhang, P., Simonsen, J.: Encoding word order in complex embeddings (2019).
21. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 1243–1252. JMLR (2017)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
23. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling (2014). arXiv:1412.3555
24. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation (2014). arXiv:1406.1078
25. Yin, W., Kann, K., Yu, M., Schütze, H.: Comparative study of cnn and rnn for natural language processing (2017). arXiv:1702.01923
26. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, June 2014
27. Kim, Y.: Convolutional neural networks for sentence classification (2014). arXiv:1408.5882
28. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology—vol. 1, NAACL '03, pp. 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics (2003)

29. Kalchbrenner, N., Blunsom, P.: Recurrent continuous translation models. In: Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1700–1709. Association for Computational Linguistics (2013)
30. Sutskever, I., Vinyals, O., Le, Q.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems (NIPS 2014)
31. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder–decoder approaches. In: Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (2014)
32. Bahdanau, D. et al.: Neural machine translation by jointly learning to align and translate. CoRR (2015). abs/1409.0473
33. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: the sparsely-gated mixture-of-experts layer (2017). arXiv:1701.06538
34. Kuchaiev, O., Ginsburg, B.: Factorization tricks for LSTM networks (2017). arXiv:1703.10722
35. Vaswani, A. et al.: Attention is all you need (2017). ArXiv abs/1706.03762
36. Devlin, J. et al.: BERT: pre-training of deep bidirectional transformers for language understanding. NAACL (2019)
37. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding with unsupervised learning. Technical report, OpenAI (2018)
38. Taylor, W.L.: Cloze procedure: a new tool for measuring readability. J. Bull. **30**(4), 415–433 (1953)
39. Jernite, Y., Bowman, S.R., Sontag, D.: Discourse-based objectives for fast unsupervised sentence representation learning. CoRR (2017). abs/1705.00557
40. Kitaev, N. et al.: Reformer: The Efficient Transformer (2020). ArXiv abs/2001.04451 (2020)
41. Lan, Z. et al.: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations (2020). ArXiv abs/1909.11942
42. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding (2019). arXiv:1906.08237
43. Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks (2015). arXiv:1511.06732
44. Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., Jurafsky, D.: Deep reinforcement learning for dialogue generation (2016). arXiv:1606.01541
45. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2013). arXiv:1312.6114
46. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
47. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space (2015). arXiv:1511.06349
48. Zhang, Y., Gan, Z., Carin, L.: Generating text via adversarial training. In: NIPS Workshop on Adversarial Training (2016)
49. Yu, L., Zhang, W., Wang, J., Yu, Y.: Seqgan: sequence generative adversarial nets with policy gradient. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)

AI and IoT Enabled Smart Hospital Management Systems

Mahendra Kumar Gourisaria, Rakshit Agrawal, Vinayak Singh, Siddharth Swarup Rautaray, and Manjusha Pandey

Abstract With the rapid increase in the advancement of different methods and technologies, efficient and optimized solutions have been introduced for providing support in the healthcare sector. With the steady increase in the population, smart healthcare system requires optimized data management algorithms for storing various data that is collected from various applications and sensors. For obtaining these objectives, integration of such modern techniques including artificial intelligence, the Internet of Things, machine learning, etc. becomes a mandatory step in developing smart hospital-based services. Artificial intelligence-based robots for surgery and diagnoses of various medical imaging are providing better results over time for smart hospital systems. IoT-based temperature management systems, sensors for checking the health of instruments, and many more applications are available for the smart hospital management system. Researchers are now becoming more penchant toward the smart hospitals, cities, etc. based development domains and have proposed various architectures which contribute to the same. These modern techniques are solving major chunks of problems in a faster way by providing good results and more facilities when compared to aged techniques. In this paper, the main focus is to provide various aspects and factors of AI and IoT-based smart hospital systems and the role of AIoT in the growth of the modern world as a combination of these niche areas is giving outstanding results in the field of healthcare. We also provide various hurdles which are encountered during the development of smart hospital management systems.

M. K. Gourisaria (✉) · R. Agrawal · V. Singh · S. S. Rautaray · M. Pandey
School of Computer Engineering, KIIT Deemed to Be University, Bhubaneswar, Odisha 751024, India
e-mail: mkgourisaria2010@gmail.com

S. S. Rautaray
e-mail: siddharthfcs@kiit.ac.in

M. Pandey
e-mail: manjushafcs@kiit.ac.in

S. S. Rautaray et al. (eds.), *Data Science in Societal Applications*, Studies in Big Data 114, https://doi.org/10.1007/978-981-19-5154-1_6

1 Introduction

Artificial Intelligence and the Internet of Things (IoT) are the most popular and enhanced technology in today's society, contributing to almost all domains from medical to home security systems. Nowadays, everyone is in search of convenient methodologies and better service with proper instructions and it is the point where IoT comes into play and acts as a crucial pillar for the enhancement of today's working of smart hospitals and other workstations. The use of IoT, cloud services, AI, and ML-based techniques can be seen in every workstation all over the globe including smart hospitals and other intelligent search-based applications. The Internet of Things [1] refers to the system or objects that are fully functional and operable via the Internet. For instance, we can consider a weather monitoring system where data is generated and collected via various sensors which will be beneficial for managing thermostats of any public infrastructures, cutting emissions, and many more. Artificial intelligence [2] is also a crucial part of modern technologies as it helps in developing a thinking ability of a machine and makes it capable to think, decide and learn. AI is helpful in the sector of healthcare by providing facilities for AI-assisted robotic surgery [3], nursing assistant [4], and decision-making like diagnosis and prediction of diseases [5].

The combination of these niche areas is efficiently solving major chunks in various domains namely the medical sector, financial sector, and other industrial sectors. Interaction between homo-sapiens and machines is enhancing day by day with the help of artificial intelligence of things (AIoT) [6]. Technologies that are getting contributions from these domains are achieving better data management, protection, analysis, data gathering, and the closest step toward efficient human–machine interaction. In the field of healthcare AIoT is providing various smart facilities like improvement in accuracy of diagnosis of disease, remote patient monitoring, reduction in needs for follow-up visits, reduction in wait times in smart hospitals by tracking bed-tracking system, and identification of critical patients who are suffering from some critical diseases and disorder. The development of various advanced machines like CT scans [7], CXR scans, and maintenance of records plays a vital role in the field of healthcare. Diagnoses of various diseases like breast cancer [8], lung cancer [9], and liver disease [10] are getting easier with the help of various AI techniques. Figure 1 shows the combination of AI and IoT which results in AIoT.

Fig. 1 Combination of AI and IoT

The healthcare system in India needs a major improvement in terms of system and management. Modern techniques like artificial intelligence and the Internet of things play an important role in the field of healthcare by optimally providing modern solutions. Optimized and efficient diagnosis of various diseases like retinal disease [11], breast cancer, diabetic retinopathy [12], lungs cancer, tuberculosis [13], and many more are possible in optimal time with less error by using smart and modern methodologies. Hospital environmental section is getting contributions from the Internet of Things where various sensors are playing an important role in maintaining temperature, humidity, and air regulation. Hospital assets are getting a predictive maintenance system for proper analysis and tracking of all the devices and machines.

The modern medical-related problem is getting solved efficiently with the help of Artificial Intelligence and the Internet of Things. With advanced technologies and introduction of smart hospital management systems, medical cases can be reduced. In the current scenario of the COVID-19 smart hospital management system worked efficiently by handling a large number of cases. The upcoming of the chapter is categorized in the following manner: 2. Literature Review, 3. Contribution, 4. Evolution of Artificial Intelligence with sub-domains and applications 5. Evolution of Internet of Things with its importance and various use cases 6. Combination of Artificial Intelligence and Internet of Things with their practical usage in healthcare 7. The smart hospital management system will discuss how these modern techniques are contributing to the field of health care.

2 Literature Review

With the steady development in the medical fields and advancements in technology, researchers have been presenting their work on the concept of smart cities, hospitals, etc. With the use of AI and IoT technology, tons of information can be analyzed by AI-based algorithms and can provide accurate results saving a lot of time and effort. Researchers have been working in the field of smart medical rooms and have proposed various architectures which help in patient monitoring and other analysis which can be found very helpful for their treatment process.

The study presented by Alharbe et al. [14], used Fuzzy-Delphi analytical hierarchy process for the evaluation purposes of various factors including the basic usable security, considering six different important factors. The combination of Fuzzy-Delphi with the analytical hierarchy process leads to an efficient diagnosis of various factors during the development of smart hospital-based systems. Many researchers have reviewed and analyzed various advanced technologies and algorithms which are found to provide exceptionally accurate and efficient results for maintaining and managing facilities of smart hospital systems. Another study was presented by Lakhoua et al. [15] where authors reviewed and analyzed various advanced techniques which are found to be helpful in the maintenance and management of smart hospitals. The paper was more penchant toward the applications based on IoT like SADT methods, UML, GRAI, GIM, etc. With the steady increase in the use of

advanced technologies and integration of these applications in the medical sector, AI and IoT have proved to be the most efficient techniques for making resources automatic and customized depending upon the environment they are used in. The study presented by Lin et al. [16] used AHP (Analytical Hierarchy Process) combined with MCDM (Multiple Criteria Decision Making) was used for providing a better evaluating criterion for smart hospitals during COVID-19 where there were mapped with BIM-related alternatives for information regarding asset information management practices.

A novel model was proposed based on Smart Hospital Management Systems by Kunar et al. [17]. The authors proposed a model focusing on a smart hospital information management system that runs by using hybrid cloud, IoT, ML, and AI-based applications. Apart from the hospital perspective, the model provided benefits to the patients as well. The model will be self-sufficient and will act as an assistant to both patients as well as the administration sector. The model will not require manual support and will be beneficial for the medical sector in long term. Bender et al. [18] provided detailed information about smart pharmaceuticals. The authors discussed various aspects of pharmaceuticals and how they can be used for providing advanced devices which are used for respiratory diseases. The study talks about the potential impacts of smart pharmaceuticals and technology that can help in the improvements. A similar study was presented by Islam et al. [19], where authors present a novel model for healthcare systems in the IoT environment that can effectively monitor patient's movement for better treatment processes. The model will also be able to provide a customized experience for the patients. The model transmitted collected data to the main server for analysis and other authentications. For this various hardware devices were used namely heartbeat, body temperature, room temperature, CO, etc. sensors. The error percentage was limited to less than 5% for every case.

The study presented by Kumar et al. [20] proposed a novel approach for generating waveforms like Non-Orthogonal Multiple Access (NOMA), Universal Filter Multi-Carrier (UFMC), and Filter Bank Multi-Carrier (FBMC) systems. Various parameters were considered in the experiment for analyzing purposes. These parameters included power spectrum density, bit error rate, capacity, Peak to Average Power (PAPR) of advanced waveforms, and Orthogonal Frequency Division Multiplexing (OFDM) methods. All these techniques were based on the development of smart hospital systems over 5G networks. Another work was presented by Amudha et al. [21], where authors proposed a system embedded with optimized and beneficial methods to predict future events based on the observations. A new framework was presented for contributing to the smart hospital management systems with the integration of intelligent decision-making, data fusion, and prediction algorithms using machine learning concepts. The authors explained various IoT-related layers including resource, communication, application, etc. layers. The chapter also explains various protocols that are used in IoT for achieving a smooth communication system. Smart automation of health monitoring frameworks and the use of cloud services for managing databases were also introduced by the authors. Apart from the frameworks, the authors also talk about the challenges faced during the development of smart hospital systems and the use of deep learning models in analyzing the

patterns in the disease from the information provided. A similar study was presented by Afferni et al. [22] where authors presented various methodologies for developing smart hospitals. The authors pointed out various key points that affect the patients which included complex organizational workflow, lack of integrated IT infrastructure in the laboratories, and other healthcare structures. The proposed framework was more oriented toward the patients' satisfaction levels. This was achieved by the interconnection of programmable collaborative robots, 3D printers which are connected to the developing servers, AR supporting the production process, etc. In our chapter, we tend to explain various other factors that contribute to the building up of smart hospital management systems and the challenges faced in the process.

3 Contribution

With the rapid growth in the fields of AI and IoT, the concept of smart cities, hospitals, etc. has become the major area for researchers. With our discussion in the chapter on the efficient and beneficial use of combined AI and IoT techniques for the development of *Smart Hospital Management Systems*. Our objective behind the chapter is to deliver various aspects by which Smart Hospitals can provide better healthcare facilities to the patients as well as staff members. Our main goals are described in the following points.

A. Detailed discussion about the evolution of Artificial Intelligence and the Internet of Things. We also present various challenges that are faced in developing such applications and their after-effects for the end-users.
B. We explore various smart management systems which include various architectures of hospitals, cities, etc.
C. Our chapter provides a proper evaluation of the limitations of all the different AIoT-based techniques.
D. Contribution of these integrated technologies in the domain of medical fields for providing advanced and customized healthcare environments to the patients.
E. Our major focus revolves around the concept of developing smart hospital systems with proper analysis and use of various technologies including RFID tags and other hardware devices.
F. Our end goal is to find a concrete path for developing a new concept of the medical systems unlike the traditional style of providing healthcare to its patients.

4 Evolution of Artificial Intelligence

From a normal medical ecosystem to a smart and intelligent medical ecosystem, artificial intelligence played an important role in its evolution. In the eighteenth century, artificial intelligence was a myth and fiction according to people but the evolution and development of technologies helped in the enhancement of machines

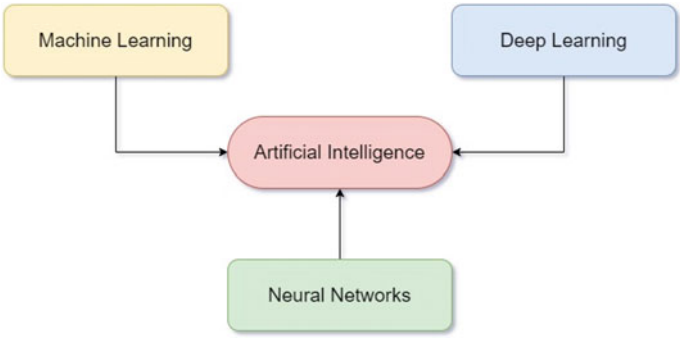


Fig. 2 Components of AI

by providing the ability to reason, making predictions, and analyze. Artificial intelligence is contributed by various sub-domains like machine learning, deep learning, and neural networks. From driverless cars to smart diagnosis of diseases is getting contribution from artificial intelligence. Artificial intelligence is the combination of 3 sub-parts, i.e., machine learning [23, 24], deep learning [25, 26], and neural networks. Figure 2 shows the formation of artificial intelligence with the help of all sub-domains.

I. Neural Networks
Neural networks are the basic component of deep learning which helps in the formation of a large network for understanding and solving large-scale problems in the healthcare system. Neural networks have proved to provide exceptional results in various domains namely fraud detection in the banking sector [27], medical sectors, etc. Artificial neural networks [28] are the basic component of neural network that is highly used for efficiently solving large-scale problems. These neural networks are based on brain-based neurons which are receiving input from one end and the output of the end works as input for other neurons. This full architecture leads to the development of fully connected layers which train them on various datasets like heart disease dataset [29], cholera dataset [30], and many more. With this advanced technique hospital management system is getting an upper edge in the diagnosis of various diseases. More accurate results are accomplished with the help of neural networks and which also results in low computational costs (Fig. 3).

II. Deep learning
Deep learning is a sub-domain of machine learning which contributes a major role in the field of artificial intelligence [31]. Almost every sector in the field of technology is getting contributions from deep learning. The healthcare domain requires high accuracy and precise results for the evaluation of results. So deep learning is essential for modern techniques to achieve high results. Smart hospital management requires an intelligent recognition system for the diagnosis of various diseases like brain tumors [32], ulcer detection

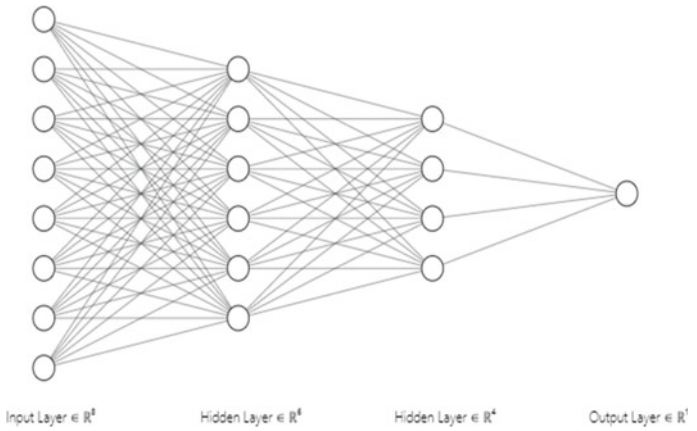


Fig. 3 The basic structure of neural networks

[33], lungs disease detection [34], ECG classifications [35], etc. Convolutional neural networks, recurrent neural networks, and transfer learning [36] are a few applications of deep learning which are helpful in the efficient diagnosis of various chronic diseases. Table 1 enlists various applications which require deep learning methods for the diagnosis.

III. Machine Learning

Machine learning is an important sub-domain of artificial intelligence. This domain enables the ability of the machine to recognize and analyze the various type of use cases. By the usage of various medical imaging techniques, several models can be developed for the proper analysis of diseases. Dangerous diseases like cancer, tumor, COVID-19 [37] can be easily diagnosed and recognized by a machine learning system where the models will be trained on various datasets for distinguishing several cases. Standard machine learning algorithms like Support vector machine, Decision tree, random forest, linear regression, logistic regression is beneficial for the prediction of various critical diseases like cholera, pneumonia disease [38, 39], cervical disease [40], heart disease diagnosis [41, 42] and many more. Figure 4 shows the basic algorithms of machine learning.

With the help of artificial intelligence, growth in the medical sector is increasing rapidly where advanced facilities like the proper diagnosis of disease, AI-assisted surgery robots, advanced machines for scanning like optical coherence tomography system [43], CT scans, and chest X-ray images, they all are advance techniques for diagnosis and detection of critical diseases. The assistance of robots for surgery, medical assistance are major applications of artificial intelligence. So, for more understanding regarding the role of AI in smart healthcare management, we are

Table 1 Deep learning-based applications

Application type	Deep learning method	Application	Input data
Medical image analysis and prognosis	Deep autoencoders Convolutional neural networks Deep belief networks	3D Brain reconstruction Neural Cells classification Classification of brain tissues	MRI scan Fundu image PET scans
	Deep autoencoders Deep neural networks	Tumor detection Cell clustering Organ segmentation	Microscopy CT image X-Ray image scan
Medical informatics	Deep autoencoders Convolutional neural networks Deep belief networks	Disease prediction Human behavior analysis Motion monitoring	HER sensor data Medical datasets Lab test reports/blood group
Public health	Deep autoencoders Convolutional neural networks Deep belief networks Deep neural networks	Predicting demographic information Lifestyle diseases Infection disease epidemics Air pollution prediction	Social Media data collection Metadata generated from mobile phones Geo-tagged images Test messages

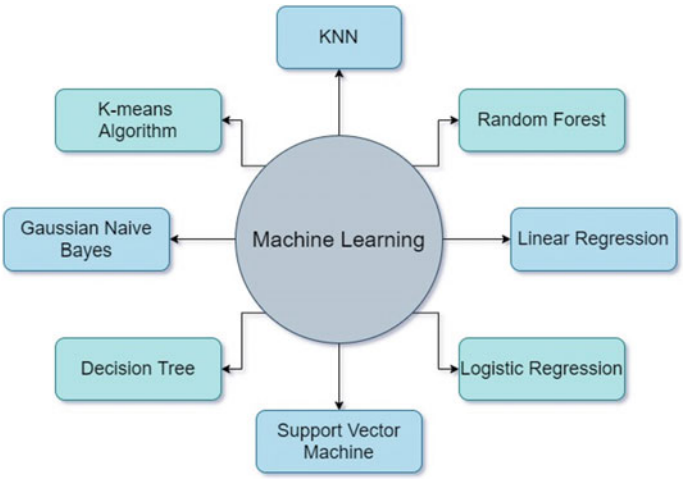


Fig. 4 Basic algorithms of machine learning

going to discuss various applications of AI in the field of health care. So, there are many applications of artificial intelligence in the medical field and a few of them have been discussed below.

A. Better diagnosis system
Healthcare requires high accuracy and precise results for the analysis of diseases. Earlier without these technologies diagnosis was time-consuming but with the help of artificial intelligence, it is a much more convenient and faster way to diagnose various chronic [44] and critical diseases. Using high-end machines for the detection of disease enables us to early diagnose of disease which will help in the reduction of repercussions. From chest X-ray to optical coherence tomography requires advanced machines for analysis of the disability. So, after reports, this medical imaging can be used for the classification and identification of patients with and without the disease [45]. With this early diagnosis, people can have their proper treatment at various medical centers. Artificial Intelligence will diagnose early but will also reduce computational costs and will save time.

B. AI-Based Surgery Robot
With the increase in the number of patients and less availability of doctors, AI-based surgery robots will play a crucial role in the medical field. These robots are highly trained in various surgery techniques like stitching, cutting, and many more. These operations can be easily performed by these AI-based robots who are under usage of reinforcement learning which analyzes and performs the best operation on various use cases according to the condition [46]. These robots are beneficial as normal operations and surgeries can be time-consuming but with the help of these technologies, we can save time with proper and fast surgeries.

C. Nurse Assistance
Artificial intelligence provides us the facility of assistance that is specific to the healthcare system [47]. With the help of this assistant people can easily have their basic treatment at home only. These assistants are trained with the help of components of machine learning, i.e., Natural Language Processing. These assistants can easily inform you about all the details about basic medication and services. So with the help of this assistance time consumption will be reduced and crowd gathering will be less. So with the help of these advanced methods smart hospital management system.

5 Challenges Faced by Artificial Intelligence

Many challenges are faced by artificial intelligence in today's real-world problems and a few of them are listed below:

- High Computing Power—Artificial Intelligence requires a high computer station for carrying out operations. Advance CPU/ GPU is required for high performance. Researchers/ workers need to maintain that computational cost is low for a given task.

- Deficiency of Trust—Major reason behind worry is due to an unknown pattern of working of the AI system. Many researchers find difficulty in finding the working pattern of AI devices.
- Less Knowledge—many peoples don't know the original potential of artificial intelligence.
- Higher accuracy with Human-Level Analysis—Machines can achieve the highest accuracy early at 95–99% but human analysis can be fully correct when compared with machines.
- Data Privacy—Data used are public and they can be used for various harmful purposes.
- Biasing Problem—The dependency of a machine is fully on the trained dataset. Dataset needs to be balanced for proper working and analysis via any workstation.

6 Evolution of Internet of Things

Internet of Things is a major component of modern technologies starting from smartphones to smart toothbrushes. An ecosystem where all devices are interconnected and transportation of data through a wired or wireless network leads to a smart IoT system. Figure 5 shows the basic structure and conversion of data for IoT systems. These generated datasets are further processed and to a cloud environment where various instances and insights are observed and decision-making takes place. From small areas like smart homes [48], smart toilets to large areas like smart hospitals [49], smart malls, and many more are focusing on an IoT-based ecosystem.

These smart devices are an important part of our lives where they help in decision-making by providing various suggestions and recommendations. These recommendations are beneficial as they help in giving an optimal solution for efficiently solving problems. Internet of Things is providing benefits on a large scale as most of the infrastructures and big giant companies are concentrating on the benefits of customers

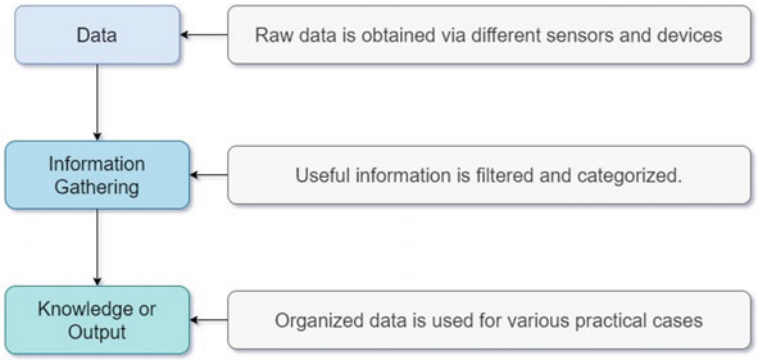


Fig. 5 Basic structural and conversion of data for IoT system

so IoT is a convenient tool for building up a better and healthy connection between servers and clients. Though IoT is not been an early bird since 1982 IoT is a crucial part of homo-sapiens. ARPANET [50] was the first connected network that used IoT and we can also call him the grandfather of the Internet and this led to the initial growth of the Internet of Things.

Internet of things is having a wide range of practical applications like.

(a) In-home—

a. Smart Lighting [51] for saving energy by adapting conditions and auto function of switching on/off according to need.
b. Smart Appliances like smart television, refrigerators, music systems, and washers/ dryers, etc. Smart thermostats for controlling the room temperature.
c. Intrusion Detection [52] is like security cameras and sensors to detect abnormal activities and raise alerts according to situations.
d. Smoke/Gas Detectors [53] are installed in the home for detection of leakage of gas or detection of fire

(b) In Industry—

a. Machine diagnosis and Prognosis for analyses of the performance of a machine by the implementation of various operations on data for estimation of various use cases under different conditions.
b. Indoor air quality monitoring [54] in factories is important for the health and safety of workers.
c. Indoor temperature monitoring system [55] for maintaining and monitoring the temperature of working areas.

(c) In Energy—

a. Smart Grids are data communication networks integrated with the electrical grid for collecting and analyzing data related to plants.
b. Renewable Energy Systems for detection of an imbalance in the ecosystem by maintaining the proper supply.
c. Prognostics for verification of all the machinery parts whether they are working inappropriate manner or not [56].

(d) In Retail—

a. Inventory Management [57] for analyses of understocking and risk of expenses and loss according to various conditions
b. Smart Payments like Paytm, BHIM for contactless payment and transferring sums of money and bounties virtually
c. Smart Vending Machines [58] is responsible for the management of inventory and providing elastic pricing of products.

(e) In Logistics—

a. Route generation and scheduling system for generation of the dataset and collecting from various sources that will be beneficial for upcoming consumers.
b. Fleet tracking [59] is used for tracking the vehicles which are used as logistic delivery vehicles by the usage of GPS technology.
c. Shipment monitoring allows having a diagnostic look according to the condition of the transportation system
d. Remote vehicle diagnostics [60] is used for the detection of a fault in delivery vehicles

(f) In agriculture—

a. Smart irrigation system [61] for watering the crops by saving water and supplying the required water at a stipulated time.
b. Greenhouse control for monitoring the condition and performance of growth of plants in greenhouses.

(g) In environment—

a. Weather Monitoring [62] system for gathering information regarding climatic change and predicting the weather of upcoming days
b. Air Pollution monitoring system [63] monitors the quality of air present in the atmosphere by comparing the percentage of all the gases present
c. Noise pollution monitoring for checking and diagnosing the frequency and decibel value of sound and categorizing it into various categories.
d. Forest Fire monitoring [64] enables to have an alert on increment and decrement of temperature which can lead to fire forest and sensors for analyzing the fire.

So, from the above points, we can notice that the Internet of Things is widely used and it is one of the flexible technologies which can be easily used and integrated with other niche areas like machine learning, deep learning, and artificial intelligence. Figure 6 shows the characteristics of IoT.

So, from all of the above discussion, we can say that the Internet of Things plays an important role in day-to-day life as it is contributing to almost every domain. Internet of Things (IoT) is helping a lot in the healthcare system by contributing to the smart hospital system. By using various monitoring systems in smart hospitals IoT provides a facility to track all the information regarding the patient's health by using various sensors for checking conditions. IoT enables us to track the health of all instruments used in surgery and are helpful for medical purpose. By usage of IoT on a large-scale healthcare is getting an edge over old techniques. Structural health monitoring of the environment of the hospital is also an application of IoT. Figure 7 shows the evolution of IoT.

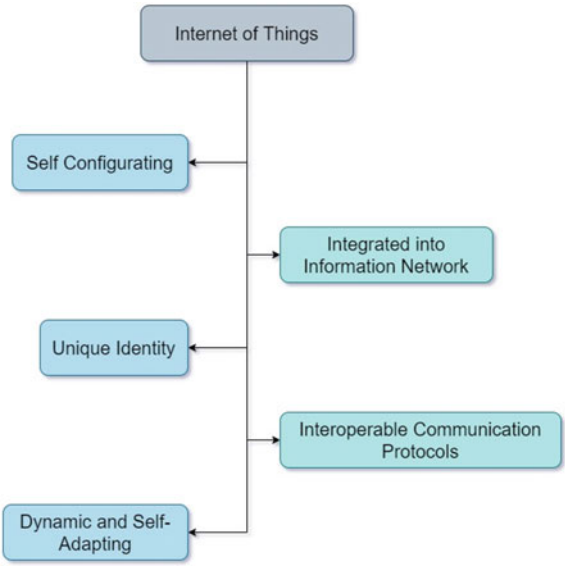


Fig. 6 Characteristics of Internet of Things

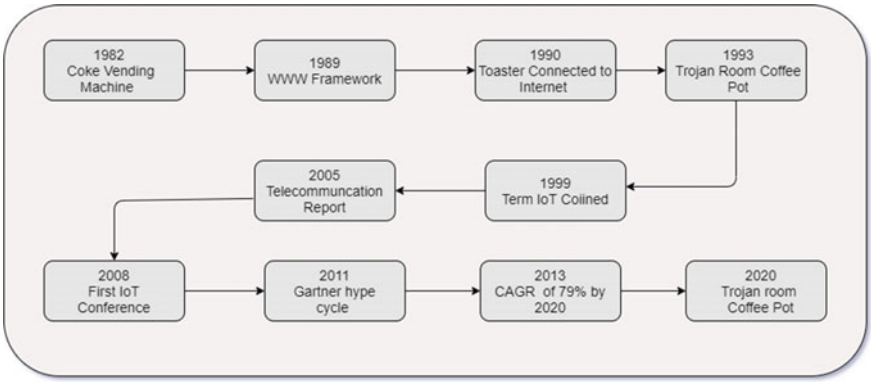


Fig. 7 Evolution of Internet of Things

Internet of Things has rapidly made a change in the lifestyle of human beings. From early 1982 till 2020 Internet of things has contributed and enhanced various domains and especially the healthcare system has received major up-gradation. So the Internet of Things has various interfaces for connecting with other devices where they can be either wired or wireless.

a. Connectivity Interface—USB Host, RJ45/ Ethernet
b. Processor/Graphics—CPU/GPU
c. Audio/Video Interface—HDMI, 3.5 mm audio, RCA

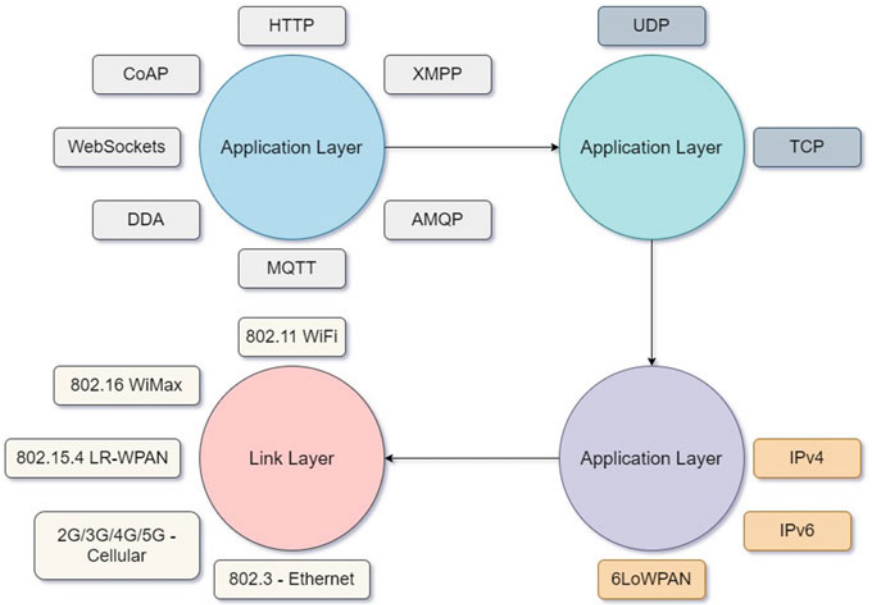


Fig. 8 Protocols of Internet of Things

d. Memory Interface—NAND/NOR, DDR1/DDR2
e. Storage Interface—SSD, MMC, HDD, SD
f. Input/ Output Interface—UART, I2C, CAN

These are the generic part of the interfaces of the Internet of Things. IoT is a vast field which is having a wide range of applications. Smart hospitals are having a smart analysis feature for tracking various instruments, health check-ups on all the instruments.

IoT helps us by keeping the track record of various medicines by having a detailed record from their manufacturing date till their expiry date which helps in marinating a hygienic and smooth operating of smart hospitals. Figure 8 shows the IoT protocols.

7 Challenges Faced by the Internet of Things

Internet of things is a vast field so it also faces some challenges which are quite important for providing better facilities. A few of the challenges are listed below:

1. Security Issue—Generation of large data expands the chance of cyber-attack and data leaks.
2. IoT regulations—Standard and quality data collection at regular intervals is necessary.

3. Compatibility—The compatibility of IoT devices requires regular updates and patch notes.
4. Bandwidth Limit—IoT requires Internet every moment for updating and monitoring data.

8 Smart Hospital Management System

With the rapid development in the medical industry and technology, patient care and satisfaction have reached new limits. The concept of SHMS comes from the optimization of file management and automation of various functionalities required by the authorities. The objective is to improve the traditional methods of storing data, treatment process, and manual analysis of various reports. Now, apart from the proper hygiene and other basic requirements, patients are more penchant toward the easiness and comfort while diagnosis and other prognosis steps. These satisfactory levels are accomplished by integrating various services namely cloud-based, IoT, Machine Learning, and AI into their management system. It is predicted that by the end of 2024, the smart hospitals market will be worth approximately 63 million dollars [65]. The start of the *Industrial Revolution 4.0* [66], brings up various AI and ML-based algorithms that help in accurately diagnosing various diseases [67], analyzing details collected from different sources. With this integration, now hospitals are customized in a more ingenious manner and personalized solutions which provide a soothing effect on the customers and hence provide a better approach toward the medical treatment of the patients. Various applications have been integrated with hospitals for making smart systems. Figure 9 illustrates the pillars of the smart hospital systems.

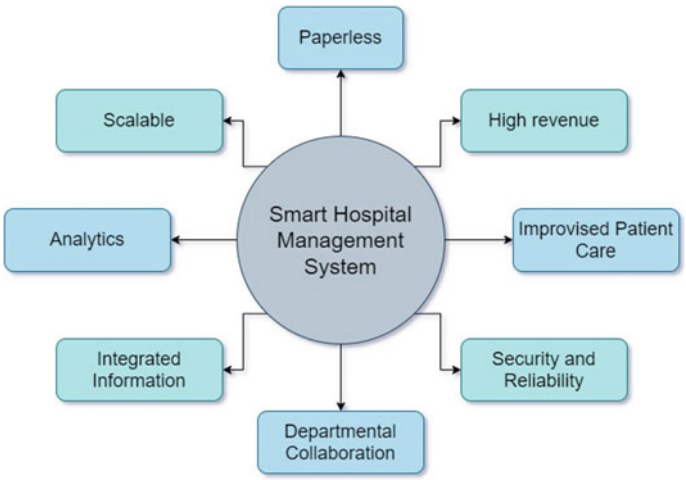


Fig. 9 Pillars of smart hospital systems

In the upcoming chapter, we present various applications involved with the Smart Hospital Management Systems (SHMS). With the integration of IoT in the management system, hospitals are now able to overcome the problem of storing limited data, inflexible networking mode, etc. The upcoming chapter is categorized in the following subsections: IoT Based Technology, Hardware Devices, Architecture model, Applications based upon smart hospital management systems, and limitations of SHMS.

I. IoT Based Technology
With the recent development, many IoT technologies have contributed to building strong information networks based on the type of integration. Following are some of the widely used techniques.

A. Internet-based Technology
With the help of the internet, the next-generation network is established within the premises of the hospital and allows the authorities to share data with other branches or the desired persons without any effort. Internet becomes the first and key technology that is required for the development of SHMS. By such integration, both patients and doctors can have friendly communication with minimum effort requirements.

B. RFID Technology
Termed as Radio Frequency Identification is a widely used technology in the healthcare sector for providing better reliability, and secure services [68]. This technology is used for the streamlined flow of heavy modules and provides full automation in the hospitals for providing details in a very optimized manner to the patients, staff, and doctors. These details include patient identification, staff allocation, doctors, medicines in the inventory, treatment details, and other necessary desired details. RFID relies upon wireless sensor-based technology which makes use of radiofrequency in forming the electromagnetic fields for transferring the data among themselves. These tags contain information stored electrically for transferring data on a large-scale basis. However, for short-range tags, magnetic fields are used and also act as a transponder for emitting microwaves. Fig. 10 demonstrates the basic components of the RFID system.

RFID has furthermore applications which are mainly classified into two categories namely *Inventory Management and Control* and *Workflow and Process Optimization.* Apart from these, RFID comes with its advantages which include patient safety (Diet, dosage monitoring), reduced medical errors (Patient identification, readable documentation), temporal data management (Hospital data on daily basis, Bi-Temporal support), provides real-time data access facility, its time and cost-effective (reduced workload on staff, reduced theft and waste), and improvised medical processes (customized billing process, patient registration process).

C. Sensor Network Technology
Various sensors are merged for building a fully functional system. Sensors being the core component of IoT, is integrated with RFID technology for

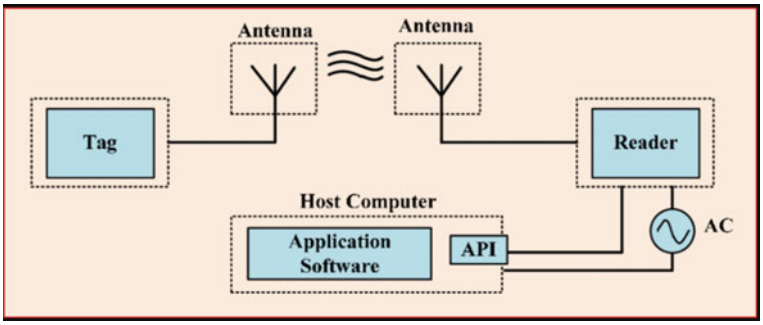


Fig. 10 RFID-based system components

better tracking and data transfer. These sensors help in tracking the various status of a variety of things including location, temperature, movements, etc. With the development of these sensors [69], making them microminiature, improved intellectual power, and collection of information, are now used in various scenarios including environmental monitoring, e-health devices, smart transporting systems, industrial plants, and military.

D. Wireless Communication and Embedded Technology
With the help of wireless connections, communication between various RFID tags can be made possible and effective. Through this network, information can be transmitted to central information systems. These wireless technologies include Bluetooth, WIFI, UWB, IrDA, etc.

II. Hardware Devices
To provide fully functional systems, various hardware units are integrated into the system that provides various information about the surroundings based upon the type of the device. Some of the widely used components are described under this sub-section.

A. ESP32 Processor
Such processors are widely used as learning tools in IoT offering Linux-based systems on scalable platforms at a minimal cost. Such processors connect sensors and actuators using GPIO pins. ESP32 provides an ingenious and customized way of developing healthcare systems. These processors contain inbuilt antenna switches, RF-balun, control amplification, low noise amplifier, various filters, and a power management system. Although these processors are combined with sensors for customized outputs, they can also work independently and ultimately reduce the overhead interaction within the main application processor. Fig. 11 demonstrates the ESP32 processor.

B. Heart Beat Sensor
These sensors are developed based on the plethysmography theory. Sensors measure the blood volume change by organs that allows a certain light

Fig. 11 ESP32 processor

intensity to pass through it. However, pulse timing becomes a critical task in the system in which the heart pulse rate is to be tracked. This tracking of heart rate tracking helps in determining blood volume. Fig. 12 shows the basic hardware used for the heartbeat sensor.

C. Body Temperature Sensor
LM35 series of the hardware device is used for accurate detection of temperature change with the output voltage. These sensors are required to be placed in the patient's body cavity for measuring the body temperature, and the data is further transmitted or collected via suitable methods. Once

Fig. 12 Heartbeat sensor

the data is collected, the data analysis process starts, and corresponding to the results proper medications are prescribed. With the data collected, some predictions can also be derived using proper AI and LM techniques. Fig. 13 demonstrates the basic LM35 sensor.

D. Room Temperature Sensor
For analysis of the room temperature, DHT11 sensors are integrated with the RFID tags. These sensors also provide information about the humidity level in the surroundings. The sensors have integrated NTC and 8-bit microcontroller chips for temperature measurement. These sensors are manufactured with such a design so that they can interconnect with other microcontrollers as well. Fig. 14 shows the room temperature sensor.

E. CO and CO_2 Sensor
MQ-9 and MQ-135 sensors are used in CO and CO_2 detection, respectively. However, MQ-9 is preferred for the detection of LPG, CO, and methane gas

Fig. 13 LM35 sensor (body temperature sensor)

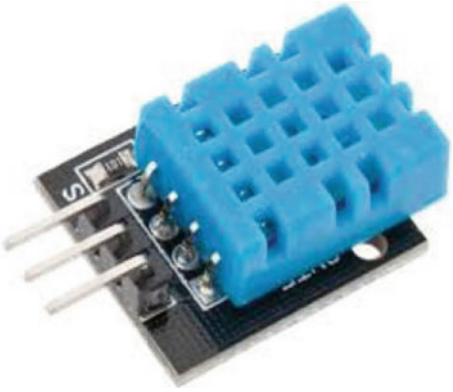


Fig. 14 Room temperature sensor

whereas MQ-135 is used for NH_3, Nicotine, Benzene, smoke, and carbon dioxide gas levels. The amount of gases is calculated in terms of PPM using analog pins and their sensitivity can be changed using the potentiometer. These sensors fit with the modern microcontroller and are widely used in the industrial domain.

III. Architecture

Based on the various proposed architectures and current requirements of the hospital systems, a customized architecture has been introduced for smart hospital systems. Fig. 15 demonstrates the block diagram of the SHMS architecture. The architecture comprises mainly three layers which include the Perception layer, Network layer, ad Application layer. These layers are described in detail in the upcoming chapter.

A. Application Layer

This layer is further bifurcated into two parts namely *Management Decision* and *Information Application.* Hospital information application includes

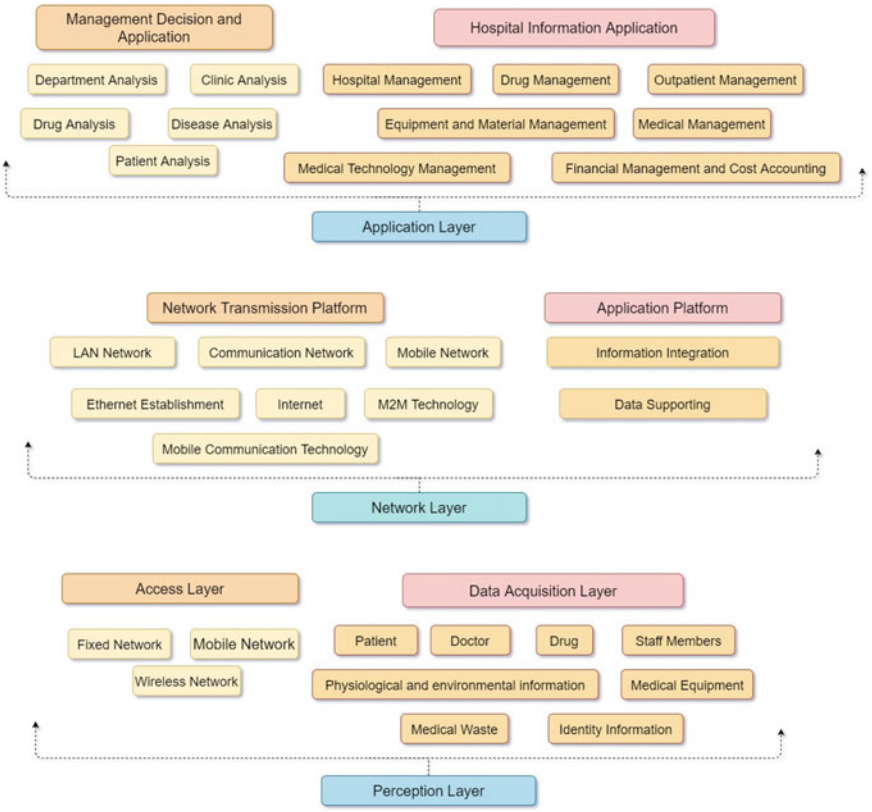


Fig. 15 Basic architectural layers for SHMS

information about the patient, hospital, medical, drug, physical therapy and materials management, medical technology, etc. On the other hand, management decision and information applications come under the senior category which includes various analyses using AI technologies. These analyses include disease (distribution based on location, expected cost of treatment, onset time), patient (age distribution, regional distribution, time of meeting, free medical services), clinic (patients visit, doctors visit), drug (amount, quantity of prescribed consumption, revenue from drugs, after-effects), etc.

B. Network Layer
The network layer has furthermore sub-layers namely *Network Transmission Platform (NTP)* and *Application Platform (AP)*. NTP becomes the backbone for the hospital networks providing real-time reliable transmission of data collected from the perception layer using various technologies (Ethernet, mobile networks, LAN, M2M, etc.). AP layer helps in the integration of various information collected to the third party for the development of the client-side interface. These client-side platforms are generally developed for the staff members of the hospitals.

C. Perception Layer
The perception layer is composed of two sub-layers which are *Data Collection* and *Access Layer*. Data Acquisition Layer is used to get a hold of hospital networking nodes, collection of surrounding data which can include information about the identity of the patient, doctors, nurse, etc., medicinal information, pharmaceutical related information, location details, etc. While Access Layer is required for transmitting the data which is collected from the other layer to the application layer or to the main server where the data can be processed and analyzed. For transmission of data, various ways can be explored including mobile networks, wireless connections, etc. These connections will become the entrance point to smart hospital systems because of their wide coverage, optimized working, cost-effectiveness, user-friendly, convenient deployment, and scalable features. Fig. 16 illustrates the basic working of the *Smart Hospital Management Systems* in a real-life scenario [19].

IV. Applications based on SHMS
In this sub-section, we provide various applications which are provided by SHM systems. These applications are acquired in the hospitals for providing a better experience to the incoming patients, doctors, staff members, and other authorities. One of the major advantages of using AIoT-based systems is the automated data management and collections which efficiently saves an immense amount of time.

A. Patient Experience
Smart hospital management systems come with customized room features for the patients and their guardians. Patient's experience can be improvised

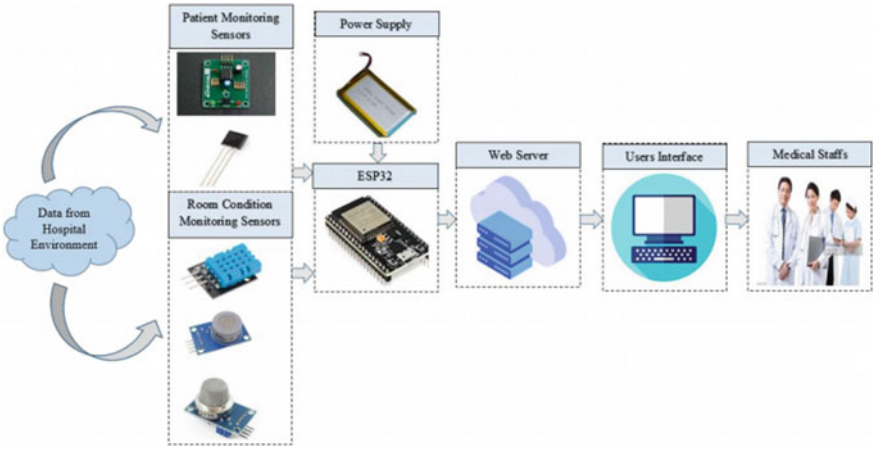


Fig. 16 Real-life smart hospital management system workflow

in various ways including remote health monitoring systems, personalized treatment, patient engagement, notification about the results, etc.

Through a remote health monitoring system, staff members can easily keep a check on the activities and other requirements of the patient from a private environment. For this IoT-based technologies play a key role, furthermore, the help of AI and ML techniques can reduce the work pressure of doctors. For instance, with the camera-based sensors and AI and ML detection algorithm models, patient movements can be viewed and can raise appropriate alarms based on them. Secondly, temperature monitoring sensors with ML techniques can help in automating the room temperature based on the required treatment process. All these sensors can be processed by the proper doctors and can watch their patients without any movement. On the same side, patients will also be able to get a personalized treatment process which will keep them motivated and joyful all way long. With medical data collection and monitoring systems in the patient's EHR, AI algorithms can provide doctors, lab staff members, and other related staff members to get end-to-end visibility of the patient's health details. These details can be found helpful while analyzing the patterns, detecting anomalies, and so on.

With the implementation of proper networks in the premises, guardians can be provided with up-to-date details and notifications about the patient's treatment process on a touch screen, or any monitor device which will improve their experience as well as patient-doctor relations. Guardians now can be involved in the care process and can have a better insight into the treatment schedules.

B. Streamlining Communication

One of the major limitations of the traditional hospital system was the improper communication between the patient and the doctors. This sometimes results in miscommunication and patients found the treatment to have overdosed. These miscommunications arise because doctors are more focused on the treatment process and other body movements which show their current state of health, while on the other hand, patients find this as doctors are ignoring them. This leads to forming a bad impression on both patients and doctors. With the integration of proper IoT technologies, patients can view their treatment details and improves timely communication between the caretakers and the patients enhancing their satisfactory level.

C. Optimized Workflow

IoT and proper network establishment can help in providing a new shape to the traditional hospital systems. Effective tools with proper functionality can provide tracking and identification services which can help in the navigation of staff members, unlike the traditional method where you have to manually go and ask for directions to meet the desired doctor. These navigation systems can also provide the live status of physicians and other staff members that are present in the hospital. Other than navigation purposes, messaging and routing features can be used to provide turn-by-turn directions and arrival time directions to make the incoming visitor less stressed about the navigation in large-scale hospitals.

D. Operational Efficiency

Operational efficiency is dependent upon two major factors namely facility management and predictivity maintenance. Facility management collects various details like temperature, humidity level, air regulation in the surroundings, security needs, etc. These details are analyzed and with the help of AIoT, efficient management systems are developed which provide automation of the actions required for maintaining proper levels of various factors (temperature, humidity, air regulation, etc.). The sensors can also help in the detection of the physical safety of the people around. With the evolution of mankind, various diseases have emerged out and will continue with more variations and effects. For achieving proper automation in the diagnosis of such variants, predictive assets of the hospitals which contribute to the healthcare domain should be updated at all time. With this, proper treatment can be given to the patients achieving better results.

E. Clinical Tasks

With the collection of real-time information from IoT-based healthcare devices, accurate diagnoses can be made by the practitioners in the hospital environment. These details can help in analyzing the illness patterns, reduce human errors, and can provide new insights with proper visualizations.

With the rapid growth in AI robots in the industries, the use of such robots is increasingly acquired by the medical industry for analyzing the hospital environment and maintaining the proper balance of various factors. These robots are also used as a helping hand in surgeries by providing reliable

instructions based on the current situation of the patient. AI-based robots contribute to providing satisfactory care to patients.

As per the survey, pharmaceutical industries spend approximately 2.7 billion dollars for every drug that comes to the market. AI algorithms can contribute to supporting drug design with an accurate prediction of molecular dynamic therapy, improving the developmental efficiency and resulting in reduced costs. Apart from drug support from AI and IoT sensors, patients can also get the daily physical therapy sessionals of their orthopedic care on mobile devices. The application can guide them through the daily exercise routine, and can also connect to the desired doctors for better communication. This will provide flexible work workflow to the patients as well as doctors who can record their respective sessions and post them on the applications as per their comfort. Also, patient's movements during the sessions can be recorded and can be analyzed by the clinicians for guiding the patients.

F. Hospital Assets Tracking
IoT-based devices are mainly used in industries for providing effective and optimized inventory tracking and pointing out the required assets beforehand. Hospitals can integrate these applications into their system for managing medicinal stocks and other requirements of the types of equipment. The application can also be used from tracking the deliveries to enable users to purchase online using preferred networks. Hospitals can make sure that the required drugs are not misplaced in between the transportation process assuring reliability to the end-users. This application, if integrated with the main server, can help in analyzing the revenue pattern of the hospital and can take further actions based on it.

G. Leveraging Data Analytics
IoT sensors can provide biometric devices which can play a vital role in the security systems of the hospital. Smart devices like wearables and smartphones can track huge volumes of biometric data. Apart from the security checks, these biometrics can be used for analysis purposes where doctors can find more about the patient as well as the population. Analysis can provide insights into the quality of patient care and can boost up medical and medicinal research.

V. Limitations of SHMS
For the fully functional development of Smart Hospital Management Systems, various hurdles are discovered by the researchers and other authorities. These challenges are described in the upcoming literature of this section.

A. Patient Monitoring
Providing monitoring services to the clinical staff members for a better treatment process becomes one of the difficult tasks. Continuous monitoring of the patient, providing satisfactory healthcare, maintaining patient safety and proper regulatory compliance with minimum staff members, having a better relationship between doctors and patients becomes the challenging

part of smart hospital management systems. To achieve the above points, a proper hold on the data analysis is required which will raise the appropriate alarms based upon the patient movements, round-the-clock care system by the doctors and staff members, efficient and fast-decision-making responses from the system, and optimized response systems are required. Developing such algorithms and sensors which can handle the above tasks, becomes a major concern for developing patient monitoring systems.

B. Data Accuracy
Data collection plays a crucial role in analyzing various patterns and the detection of diseases. Correct transmission of data from the nodes to the main server becomes a challenging part. This is indeed a difficult task because in case if wrong information or data is used for the detection in the AI and ML-based models, false predictions can be generated creating very severe side effects for the patients to suffer from. Proper collection of data and effective transmission of the same becomes the key to the smart hospital management systems and to make them fully automated ones. For ensuring that the collected data is accurate, EHR systems are used which helps in rectifying the same problem. However, there is still a scope of human error that can become a hurdle in some situations.

C. Security and Privacy
The data collected about the patient, staff, doctor's identity, treatment process, ongoing projects in the hospitals, healthcare data, inventory details, etc. should be used over a secure network. Maintaining such confidentiality, integrity, and availability of healthcare data becomes another challenge in building up smart systems. Storing these identification data, and other treatments on unreliable servers can cause damage to the hospital as well as can make the patients vulnerable to various problems which include chances of abuse, tempering of the data, etc. [70]. A compromise in the patient's or staff member's privacy can bring up hurdles and cause of various other problematic factors as well as various other implications in real-life.

D. Cost-Effectiveness
Healthcare treatment cost becomes another constraint for patients to come and rectify their problems. As per the survey, a huge number of patients come from a middle-class background and are not able to afford the treatment costs. These treatment costs include doctors and helping staff members' fees, facilities provided, medicinal charges, etc. With the integration of smart applications in the hospital, although the cost of integration becomes a one-time investment, proper maintenance and updating have costs that add to the treatment bills. From the past decades, there has been a clash between the user convenience and cost of the treatment [71]. However, there are ways by which these cost factors can be reduced in a long-term session. With the use of reliable cloud services, start-up expenses can be reduced and can be found very efficient in handling databases. EHR becomes the main component in building up a *Smart Hospital Management* system, however, security and privacy issues come into the picture with it.

E. Intelligent Data Preprocessing and Data Validation
All the IoT-based sensors send their collected data to the main node or server where every piece of information is stored depending upon the type of it. In real-life scenarios, there are many a time when data preprocessing becomes a crucial point before providing the data to any other algorithm for analyzing process. Having a variety of sensors and different sources of data collection, healthcare data becomes very diverse and heterogeneous making it difficult for the algorithms to predict accurate results. Extraction of useful information from the data cluster and validating the information becomes a challenging part. However, with the use of advanced AI-based algorithms, this process of information extraction can be made less stressful.

F. Management
Healthcare departments deliver various options to their end-users. These services include healthcare quality, safety measures, secured data services, financial productivity, regulatory standards, etc. With all these facilities, authorities have to divide the financial cost for each domain. Because of these measures and customizations in the facilities, there are conflicts of priorities as to which domain gives more importance. However, for building smart hospitals, these facilities can be embedded in a sequential order instead of installing all the applications in a single go. While the development of these applications takes place, data collection becomes another barrier as there are unstructured data integration problems. Another challenge faced by the staff members is to get acquainted with the new environment of automated functional applications and use them in an efficient way which requires quite a time for the staff members to get a hold of it.

9 Conclusion

With the aging population and advancements in the field of technology and medicinal development, improvising healthcare results has become the main focus of doctors and researchers. With every passing day, there is a new variant of chronic disease, increasing healthcare costs, and rising demands and expectations from the patients. As per the survey, approximately 8.8 trillion dollars were spent in 2020 by global health organizations. To meet the patient's satisfactory level and provide them effective treatment becomes the main objective of the *Smart Hospital Management Systems.* Smart hospital management system brings the revolution of the patients as well for the doctors and staff members. With the integration of various modern and efficient technology for the automation systems and innovations in the IT field, smart hospitals cover various aspects of the hospital which includes a variety of operations, building design, patient experience, and many more. With the help of AI and IoT, smart hospitals are developing at a faster rate, providing patient-friendly treatments, and a more affordable diagnosis process. The concept of a smart hospital system will reduce the time constraint for the patients and will have better healthcare systems

to make use of. In this chapter, we provide various aspects of AI, IoT, applications of SHMS, and limitations. We also present the architectural block diagram which provides a basic idea about the working of the SHM systems. In our chapter, we also provide various challenges that are faced by Smart Hospitals and their respective mitigations. With the steady increase in the treatment costs and demands, smart hospitals have come out to be an efficient solution for providing more efficient diagnosis processes and a friendly environment to the patients. IoT and AI-based technology reduce the expense factor to a great extent making it feasible for the patients to take up their treatments without any further stress. With the rapid growth in technologies, namely, AI, machine learning, IoT, big data analytics, cloud services, and the development of smart hospitals systems will become smoother than ever.

References

1. Xia, F., Yang, L.T., Wang, L., Vinel, A.: Internet of things. Int. J. Commun. Syst. **25**(9), 1101 (2012)
2. Nilsson, N.J.: The Quest for Artificial Intelligence. Cambridge University Press (2009)
3. Hashimoto, D.A., Rosman, G., Rus, D., Meireles, O.R.: Artificial intelligence in surgery: promises and perils. Ann. Surg. **268**(1), 70 (2018)
4. Pineau, J., Montemerlo, M., Pollack, M., Roy, N., Thrun, S.: Towards robotic assistants in nursing homes: challenges and results. Robot. Auton. Syst. **42**(3–4), 271–281 (2003)
5. Schmidt-, U., Waldstein, S.M., Klimscha, S., Sadeghipour, A., Hu, X., Gerendas, B.S., Bogunović, H.: Prediction of individual disease conversion in early AMD using artificial intelligence. Invest. Ophthalmol. Vis. Sci. **59**(8), 3199–3208 (2018)
6. Ghosh, A., Chakraborty, D., Law, A.: Artificial intelligence in Internet of Things. CAAI Trans. Intell. Technol. **3**(4), 208–218 (2018)
7. Ramakrishnan, S., Nagarkar, K., DeGennaro, M., Srihari, M., Courtney, A. K., Emick, F.: A study of the CT scan area of a healthcare provider. In: Proceedings of the 2004 Winter Simulation Conference, vol. 2, pp. 2025–2031. IEEE (2004)
8. Rodríguez-Ruiz, A., Krupinski, E., Mordang, J.J., Schilling, K., Heywang-Köbrunner, S.H., Sechopoulos, I., Mann, R.M.: Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology **290**(2), 305–314 (2019)
9. Borrelli, P., Ly, J., Kaboteh, R., Ulén, J., Enqvist, O., Trägårdh, E., Edenbrandt, L.: AI-based detection of lung lesions in [18 F] FDG PET-CT from lung cancer patients. EJNMMI Phys. **8**(1), 1–11 (2021)
10. Chen, M., Zhang, B., Topatana, W., Cao, J., Zhu, H., Juengpanich, S., Cai, X.: Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. NPJ Precis. Oncol. **4**(1), 1–7 (2020)
11. Yang, C.H.H., Huang, J.H., Liu, F., Chiu, F.Y., Gao, M., Lyu, W., Tegner, J.: A novel hybrid machine learning model for auto-classification of retinal diseases (2018). arXiv preprint arXiv: 1806.06423
12. Priya, R., Aruna, P.: Diagnosis of diabetic retinopathy using machine learning techniques. ICTACT J. Soft Comput. **3**(4), 563–575 (2013)
13. Lakhani, P., Sundaram, B.: Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology **284**(2), 574–582 (2017)
14. Alharbe, N.: 'A fuzzy-Delphi based decision-making process for measuring usable-security of Web based smart hospital management system. ICIC Express Lett. **14**(1), 15–21 (2020)

15. Lakhoua, N.: Review on smart hospital management system technologies. Res. Sci. Today **1**, 187–194 (2019)
16. Lin, C.L., Chen, J.K., Ho, H.H.: BIM for smart hospital management during COVID-19 using MCDM. Sustainability **13**(11), 6181 (2021)
17. Kumar, J.N.A., Suresh, S.: A proposal of smart hospital management using hybrid cloud, IoT, ML, and AI. In: 2019 International Conference on Communication and Electronics Systems (ICCES), pp. 1082–1085). IEEE (2019)
18. Bender, B.G., Chrystyn, H., Vrijens, B.: Smart pharmaceuticals. In: Health 4.0: How virtualization and big data are revolutionizing healthcare, pp. 61–90). Springer, Cham (2017)
19. Islam, M.M., Rahaman, A., Islam, M.R.: Development of smart healthcare monitoring system in IoT environment. SN Comput. Sci. **1**, 1–11 (2020)
20. Kumar, A., Dhanagopal, R., Albreem, M.A., Le, D.N.: A comprehensive study on the role of advanced technologies in 5G based smart hospital. Alex. Eng. J. **60**(6), 5527–5536 (2021)
21. Amudha, S., Murali, M.: Enhancement of IoT-based smart hospital system survey paper. In: Edge Computing and Computational Intelligence Paradigms for the IoT, pp. 238–261. IGI Global (2019)
22. Afferni, P., Merone, M., Soda, P. Hospital 4.0 and its innovation in methodologies and technologies. In: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), pp. 333–338). IEEE (2018)
23. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015)
24. Khare, S., Gourisaria, M.K., Harshvardhan, G.M., Joardar, S., Singh, V.: Real estate cost estimation through data mining techniques. In: IOP Conference Series: Materials Science and Engineering, vol. 1099, No. 1, p. 012053. IOP Publishing (2021)
25. Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Asari, V.K.: A state-of-the-art survey on deep learning theory and architectures. Electronics **8**(3), 292 (2019)
26. Gourisaria, M.K., Das, S., Sharma, R., Rautaray, S.S., Pandey, M.: A deep learning model for malaria disease detection and analysis using deep convolutional neural networks. Int. J. Emerg. Technol. **11**(2), 699–704 (2020)
27. Sahu, A., Harshvardhan, G.M., Gourisaria, M.K.: A dual approach for credit card fraud detection using neural network and data mining techniques. In: 2020 IEEE 17th India Council International Conference (INDICON), pp. 1–7. IEEE (2020)
28. Wu, Y.C., Feng, J.W.: Development and application of artificial neural network. Wireless Pers. Commun. **102**(2), 1645–1656 (2018)
29. Khemphila, A., Boonjing, V.: Heart disease classification using neural network and feature selection. In: 2011 21st International Conference on Systems Engineering, pp. 406–409. IEEE (2011)
30. Pezeshki, Z., Tafazzoli-Shadpour, M., Nejadgholi, I., Mansourian, A., Rahbar, M.: Model of cholera forecasting using artificial neural network in Chabahar City, Iran. Int. J. Enteric. Pathog. **4**(1), 1–8 (2016)
31. Sun, R.: Optimization for deep learning: theory and algorithms (2019). arXiv preprint arXiv:1912.08957
32. Mohsen, H., El-, E.S.A., El-, E.S.M., Salem, A.B.M.: Classification using deep learning neural networks for brain tumors. Futur. Comput. Inform. J. **3**(1), 68–71 (2018)
33. Goyal, M., Hassanpour, S.: A refined deep learning architecture for diabetic foot ulcers detection (2020). arXiv preprint arXiv:2007.07922
34. Bhatia, S., Sinha, Y., Goel, L.: Lung cancer detection: a deep learning approach. In: Soft Computing for Problem Solving, pp. 699–705. Springer, Singapore (2019)
35. Sharma, R., Gourisaria, M.K., Rautaray, S.S., Pandey, M., Patra, S.S.: ECG classification using deep convolutional neural networks and data analysis. Int. J. Adv. Trends Comput. Sci. Eng. **9**, 5788–5795 (2020)
36. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2009)

37. Punn, N.S., Sonbhadra, S.K., Agarwal, S.: COVID-19 epidemic analysis using machine learning and deep learning algorithms. MedRxiv (2020)
38. Cooper, G.F., Aliferis, C.F., Ambrosino, R., Aronis, J., Buchanan, B.G., Caruana, R., Spirtes, P.: An evaluation of machine-learning methods for predicting pneumonia mortality. Artif. Intell. Med. **9**(2), 107–138 (1997)
39. Harshvardhan, G.M., Gourisaria, M.K., Rautaray, S.S., Pandey, M.: Pneumonia detection using CNN through chest X-ray. J. Eng. Sci. Technol. 16(1), 861–876 (2021)
40. Unlersen, M.F., Sabanci, K., Özcan, M.: Determining cervical cancer possibility by using machine learning methods. Int. J. Latest Res. Eng. Technol. **3**(12), 65–71 (2017)
41. Nayak, S., Gourisaria, M.K., Pandey, M., Rautaray, S.S.: Prediction of heart disease by mining frequent items and classification techniques. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 607–611. IEEE (2019)
42. Nayak, S., Gourisaria, M.K., Pandey, M., Rautaray, S.S.: Heart disease prediction using frequent item set mining and classification technique. Int. J. Inf. Eng. & Electron. Bus. 11(6) (2019)
43. Podoleanu, A.G., Rogers, J.A., Jackson, D.A., Dunne, S.: Three dimensional OCT images from retina and skin. Opt. Express **7**(9), 292–298 (2000)
44. Hegde, S., Mundada, M.R.: Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence based feature selection approach. Int. J. Pervasive Comput. Commun. (2020)
45. Shahbaz, M., Ali, S., Guergachi, A., Niazi, A., Umer, A.: Classification of Alzheimer's disease using machine learning techniques. In: DATA, pp. 296–303 (2019)
46. Bodenstedt, S., Wagner, M., Müller-Stich, B.P., Weitz, J., Speidel, S.: Artificial intelligence-assisted surgery: potential and challenges. Visc. Med. **36**(6), 450–455 (2020)
47. McGrow, K.: Artificial intelligence: essentials for nursing. Nursing **49**(9), 46 (2019)
48. Al-Ali, A.R., Zualkernan, I.A., Rashid, M., Gupta, R., AliKarar, M.: A smart home energy management system using IoT and big data analytics approach. IEEE Trans. Consum. Electron. **63**(4), 426–434 (2017)
49. Yu, L., Lu, Y., Zhu, X.: Smart hospital based on internet of things. J. Networks **7**(10), 1654 (2012)
50. Hauben, M.: History of ARPANET. Site de l'Instituto Superior de Engenharia do Porto 17 (2007)
51. Sikder, A.K., Acar, A., Aksu, H., Uluagac, A.S., Akkaya, K., Conti, M.: IoT-enabled smart lighting systems for smart cities. In: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), pp. 639–645. IEEE (2018)
52. Zhao, L., Matsuo, I.B.M., Zhou, Y., Lee, W.J.: Design of an industrial IoT-based monitoring system for power substations. IEEE Trans. Ind. Appl. **55**(6), 5666–5674 (2019)
53. Shah, R., Satam, P., Sayyed, M.A., Salvi, P.: Wireless smoke detector and fire alarm system. Int. Res. J. Eng. Technol. (IRJET) **6**(1), 1407–1412 (2019)
54. Gunawan, T.S., Munir, Y.M.S., Kartiwi, M., Mansor, H.: Design and implementation of portable outdoor air quality measurement system using Arduino. Int. J. Electr. Comput. Eng. **8**(1), 280 (2018)
55. Karami, M., McMorrow, G.V., Wang, L.: Continuous monitoring of indoor environmental quality using an Arduino-based data acquisition system. J. Build. Eng. **19**, 412–419 (2018)
56. Hossein, N., Mohammadrezaei, M., Hunt, J., Zakeri, B.: Internet of Things (IoT) and the energy sector. Energies **13**(2), 494 (2020)
57. Jayanth, S., Poorvi, M.B., Sunil, M.P.: Inventory management system using IOT. In: Proceedings of the First International Conference on Computational Intelligence and Informatics, pp. 201–210. Springer, Singapore (2017)
58. Cho, B.H., Ahn, H.H.: Analysis and design of smart vending machine system based on IoT. J. Inst. Internet, Broadcast. Commun. **19**(3), 121–126 (2019)
59. Penna, M., Arjun, B., Goutham, K.R., Madhaw, L.N., Sanjay, K.G. Smart fleet monitoring system using Internet of Things (IoT). In: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp. 1232–1236. IEEE (2017)

60. Dragojević, M., Stević, S., Stupar, G., Živkov, D.: Utilizing iot technologies for remote diagnostics of next generation vehicles. In: 2018 IEEE 8th International Conference on Consumer Electronics-Berlin (ICCE-Berlin), pp. 1–4. IEEE (2018)
61. Vaishali, S., Suraj, S., Vignesh, G., Dhivya, S., Udhayakumar, S.: Mobile integrated smart irrigation management and monitoring system using IOT. In: 2017 international conference on communication and signal processing (ICCSP), pp. 2164–2167. IEEE (2017)
62. Chandana, L.S., Sekhar, A.R.: Weather monitoring using wireless sensor networks based on IOT. Int. J. Sci. Res. Sci. Technol **4**, 525–531 (2018)
63. Ayele, T.W., Mehta, R.: Air pollution monitoring and prediction using IoT. In: 2018 second international conference on inventive communication and computational technologies (ICICCT), pp. 1741–1745). IEEE (2018)
64. Basu, M.T., Karthik, R., Mahitha, J., Reddy, V.L. IoT based forest fire detection system. Int. J. Eng. Technol. 7(2.7), 124–126 (2018)
65. Retrieved on 27th August 2021 from https://cprimestudios.com/blog/what-smart-hospital-and-how-build-your-own-solution
66. Gourisaria, M.K., Agrawal, R., Harshvardhan, G.M., Pandey, M., Rautaray, S.S.: Application of machine learning in industry 4.0. Mach. Learn.: Theor. Found. Pract. Appl. 57–87 (2021)
67. Gourisaria, M.K., Harshvardhan, G.M., Agrawal, R., Patra, S.S., Rautaray, S.S., Pandey, M.: Arrhythmia detection using deep belief network extracted features from ECG signals. Int. J. E-Health Med. Commun. (IJEHMC) **12**(6), 1–24 (2021)
68. Lahtela, A.: A short overview of the RFID technology in healthcare. In: 4th International Conference on Systems and Networks Communication, pp. 165–169 (2009)
69. Atzori, L., Iera, A., Morabito, G.: The internet of things: a survey. Comput. Netw. **54**(15), 2787–2805 (2010)
70. Qi, J., Yang, P., Min, G., Amft, O., Dong, F., Xu, L.: Advanced internet of things for personalized healthcare systems: a survey. Pervasive Mob. Comput. **41**, 132–149 (2017)
71. Sebastian, M.P.: Smart Healthcare: Challenges and Opportunities. International Academic Conference on Management, Economics and Marketing, July 06–07, Vienna, pp. 396–403 (2018)

Optimization in Agricultural Growth Using AI and Satellite Imagery

Bhaskar Goswami and Plaban Nayak

Abstract In the domain of agriculture, there are issues such as variations in the soil profile, changes in the climatic conditions, and crops being spoiled by weeds. These issues are being addressed by technological advancements in the field of Artificial Intelligence (AI) and Satellite Imagery. In this era of Smart Agriculture, there are many ways in which technology acts as a blessing to farmers. It can be used to pick fruits or crops, detect weeds or diseases in crops through the use of satellite or drone images. With the use of current technology, we can optimize how fertilizers are applied to maintain the health of the crops. Additionally, satellite imagery can be very useful in getting real-time weather forecasts that can be helpful to farmers. There are also prescriptive ways of using AI to aid the farmers such as analyzing soil samples and recommending the kind of treatment that can be done to improve soil quality, the kind of crops that can be grown or should be grown is very helpful to farmers across the world. Crop growth can also be monitored through drones and contribute greatly to increasing the quality of the produce, thereby adding to the quality of life of a farmer. This chapter contains an in-depth analysis of the life cycle of farming and agriculture, its issues, and how AI can help address those issues by optimizing the way farming is done in this technology era.

Keywords Artificial intelligence · Agriculture · Satellite imagery · Crop analysis · Soil profiling · Real-time weather forecasting

1 Introduction

In today's smart world where technological advancements are happening every minute and every second of our life is dependent on technology, the spread of artificial intelligence, machine learning, and data mining through satellite or drone imagery in the field of agriculture is rapid. Technology plays a vital role in the development of new agricultural practices and enhancements to the conventional methods of farming.

B. Goswami (✉) · P. Nayak
HighRadius Technologies, Hyderabad, India
e-mail: goswami.bhaskar1998@gmail.com

S. S. Rautaray et al. (eds.), *Data Science in Societal Applications*, Studies in Big Data 114, https://doi.org/10.1007/978-981-19-5154-1_7

This is popularly known as “Smart Farming”. It is the boost that farming needed to make agriculture-heavy countries the forerunner of world domination, to become the most influential nations of the world.

1.1 Artificial Intelligence and Satellite Imagery

Artificial Intelligence can be defined as a way in which computer systems or machines, in general, try to simulate human intelligence. Some applications of AI include natural language processing, speech recognition, expert systems robotics, machine learning, and machine vision as shown in Fig. 1. However, machine vision, image detection, and classification are the main applications of artificial intelligence in the scope of this chapter.

Satellite images are images of Earth collected by imaging satellites operated by various government organizations and also private businesses for their products or analysis. They are the most powerful tool used by meteorologists. They are used to capture the behavior of the atmosphere and the analysis of such satellite images is necessary for accurate and real-time weather forecasts. Drone Satellites are used to

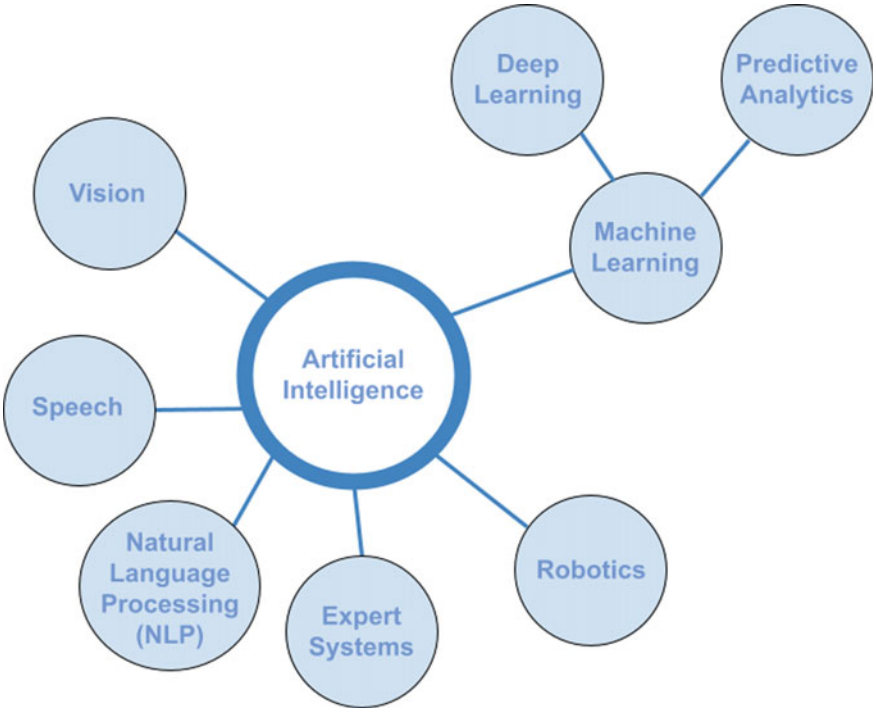


Fig. 1 Artificial intelligence and its various applications

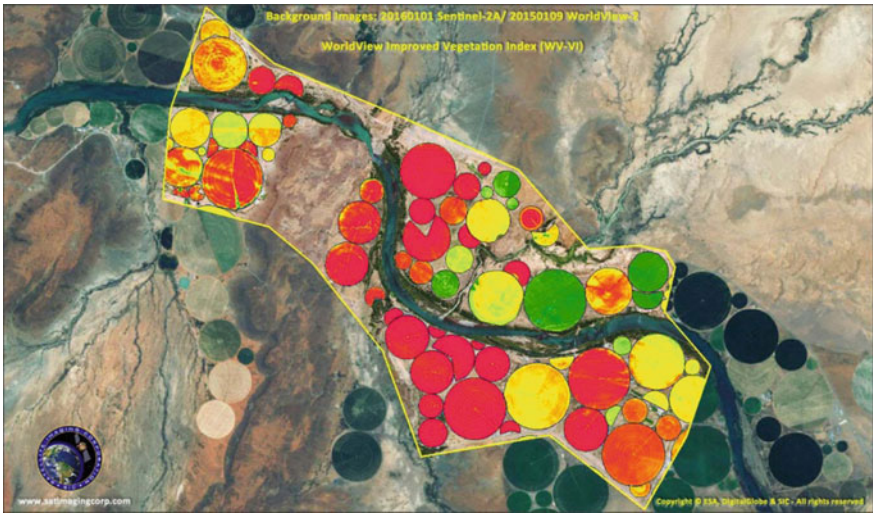


Fig. 2 Agricultural mapping of crops using satellite imaging

get data relating to a particular land. It can have applications such as weed detection, crop disease detection, and many others. They can provide reliable data about huge patches of farms and are more efficient than manual inspections [1]. The use of satellite imagery in agriculture [2] is best shown in Fig. 2. where the kind of crops grown in various areas of a huge farm is recognized and classified by AI.

1.2 Agriculture Industry: Life Cycle

Agriculture can be defined as the scientific practices of farming which include cultivation of the soil for the growing of crops and the breeding of animals to provide food wool and other products to everyone. Agriculture can be divided into various phases and has a particular lifecycle as depicted in Fig. 3.

Preparation of soil: It is the initial stage of farming where farmers prepare the soil for sowing seeds. This process involves breaking large soil clumps and remove debris, such as sticks, rocks, and roots. Additionally, fertilizers and organic matter are added depending on the type of crop to create an ideal situation for crops.

Sowing of seeds: This stage requires taking care of the distance between two seeds and the depth for planting seeds. At this stage climatic conditions such as temperature, humidity, and rainfall are also taken into consideration.

Adding Fertilizers: To maintain soil fertility is an important factor so the farmer can continue to grow nutritious and healthy crops. Farmers turn to fertilizers because these substances contain plant nutrients such as nitrogen, phosphorus, and potassium.

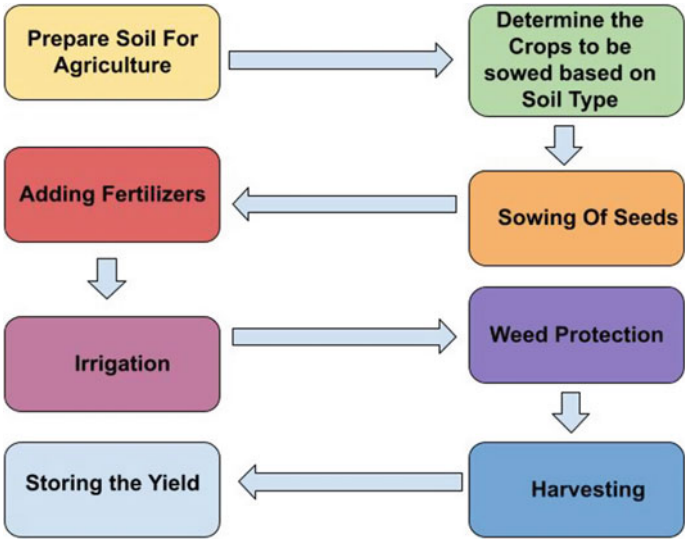


Fig. 3 Life cycle of Agriculture

Fertilizers are added nutrients applied to agricultural fields to supplement the required elements found naturally in the soil. This stage also determines the quality of the crop.

Irrigation: This stage helps to keep the soil moist and maintain humidity. Underwatering or overwatering can hamper the growth of crops and if not done properly it can lead to damaged crops.

Weed protection: Weeds are unwanted plants that grow near crops or at the boundary of farms. Weed protection is an important factor as weed decreases yields, increases production cost, interferes with harvest, and lowers crop quality.

Harvesting: It is the process of gathering ripe crops from the fields. It requires a lot of laborers for this activity so this is labor-intensive. This stage also includes post-harvest handling such as cleaning, sorting, packing, and cooling.

Storage: This phase of the post-harvest system comprises how the products are kept to guarantee food security. It also includes packing and transportation of crops along with storing them for long durations of time in a climate-sensitive manner with the help of technology.

2 Challenges Faced by Farmers in Traditional Methods of Farming

There are many challenges in the traditional methods of farming that are currently being used by hundreds of farmers in many developing countries. Some of the main challenges as depicted in Fig. 4 have been mentioned below.

Firstly, in farming climatic factors such as rainfall, temperature changes, humidity does play an important role in the agriculture lifecycle. The increase in global temperature due to global warming, pollution, and deforestation causes confusion among farmers about the timeline they should follow to prepare the soil, sow seeds, irrigate their crops, and harvest them.

Secondly, irrigation plays an important role in the growth of crops. Having knowledge about rainfall that is about to happen is a blessing to farmers so that they can regulate the amount of water that they supply to their crops. In addition to that various crops have different water requirements, and managing that is a very arduous task for the farmers.

Along with that, the requirements of the crop for various nutrients may vary. The main nutrients present in the soil are Nitrogen (N), Phosphorous (P), and Potassium (K). The deficiency of these nutrients may lead to poor quality of crops. Farmers are also not aware of the type of crops that should be grown on the soil profile of a particular place. Sometimes this results in poor harvest and is a loss to the farmer because of the lack of information.

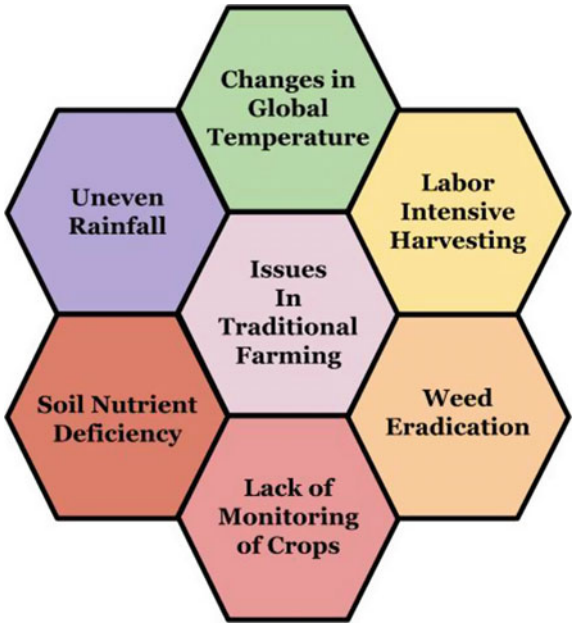


Fig. 4 Issues in traditional methods of farming

Another issue that is faced by farmers worldwide is the problem of harvesting which is a very labor-intensive task. Picking up crops, fruits and judging if they are ripe or not is a taxing and mammoth task for the farmers. It needs a trained eye to understand if the crop is ready to be harvested. Failing that may lead to the loss of huge amounts of produce and money which is not ideal for the farmer.

Weed eradication is also one of the most tricky challenges that a farmer faces. It is tricky because the use of too many pesticides may cause harm to the product and too little may cause the weed to grow and destroy the crops. The use of pesticides should be moderated and also targeted to those parts of the farm that have weed growth. This is a difficult task for the farmer who does not have the time to analyze the farm and get a holistic view of the weed growth on his land.

3 Artificial Intelligence and Satellite Imagery in Agriculture

Artificial intelligence, machine learning, satellite, and drone imagery play a vital role in the growth and development of agriculture all over the world. There are many use-cases where artificial intelligence is helping pave the way for more advanced techniques to increase the production levels in addition to making farming and the life of farmers more efficient and effective. The various issues faced by farmers in the traditional methods of farming are all being tackled by development in the field of artificial intelligence. Some of them are discussed in detail in this chapter.

3.1 Crop Yield Prediction with AI and Machine Learning

Climate and its unpredictability are some of the major factors responsible for the reduction of the rate of crop production throughout the world. Thus, weather forecasting has become essential for the proper management of crops. For instance, in industries like sugar that are heavily dependent on sugarcane farmers or any other agricultural produce suffer because of low production rates.

Crop yield prediction with the help of various complex algorithms, artificial intelligence, machine learning, and satellite imagery helps in planning and executing the logistical requirements of the business. The application of Artificial Intelligence and the use of other relevant IoT (Internet of Things) related devices are prevalent in the AgTech industry.

Crop yield prediction with various state-of-the-art algorithms, AI, and ML has many benefits for the agricultural industry and its farmers as shown in Fig. 5.

Remote Monitoring of Farm Areas is a very important benefit of crop yield prediction [3]. It provides the farmers the option to remotely monitor the yield and rate of crop growth of huge areas of land. Without the intervention of AI and technology,

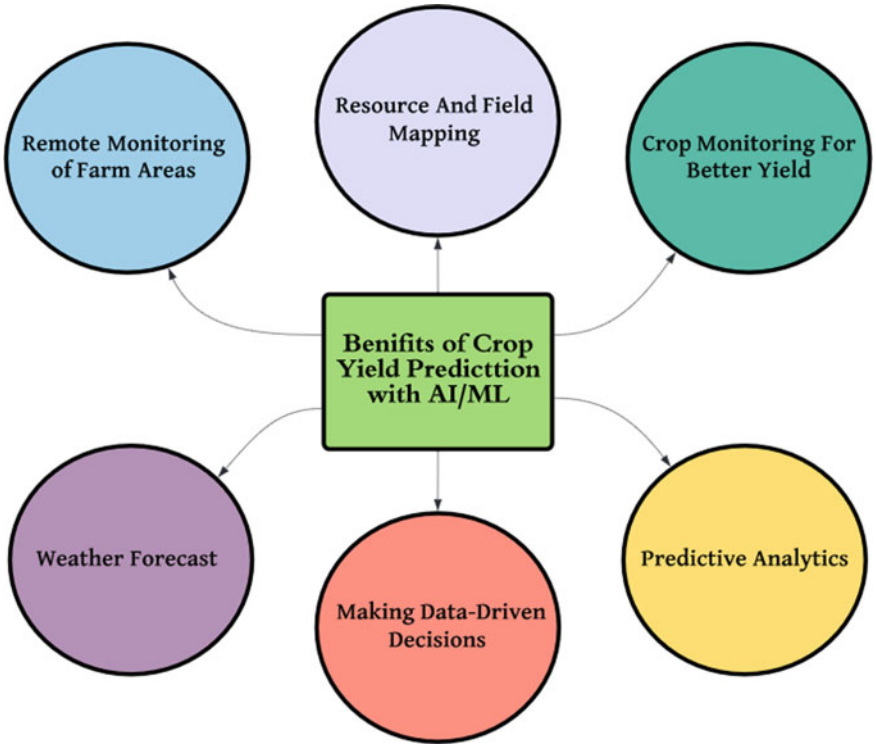


Fig. 5 Benefits of crop yield prediction

this would not have been possible. In Fig. 6 we can see the use of drone and satellite imagery to remotely monitor the growth of entire farmland that is probably acres large and is impossible for humans to monitor on their own.

Resource and Field Mapping can be very useful while tackling huge areas of farmland. It can help to understand which area of the land needs what amount of resources and allocate it accordingly. It can also help in understanding the distribution of crops that are grown on the farm. It can be helpful while harvesting to understand the priority in which the crops need to be harvested. The image below Fig. 7 shows an example of the use of satellite imagery to perform precision agriculture.

Crop Monitoring for better Yield is very useful to farmers. It helps them monitor the growth of the crops and decide the timeline for harvesting and creating a plan of how they would go about executing the harvesting process. In addition to that, it can be harnessed in detecting weeds, diseases, and pests in crops. This is helpful as it provides the farmer the information he needs to act upon and use pesticides on those areas of the farm. In Fig. 8 we can see the use of satellite imagery to monitor crops over huge areas of farmland [4, 5].

Fig. 6 Precision agriculture using satellite

Predictive Analytics and **Making Data-Driven Decisions** are harnessed in many ways in the field of agriculture. The prediction of the amount of rain in a particular season or on a daily basis can suggest to the farmers the amount of irrigation that may be needed by the farm. It can also be used to predict the amount of pesticide required for a farm keeping in mind the amount of land that has been affected by a pest or any other disease. These developments in the field of AI, ML, and satellite imagery have made it possible for technology to assist farmers to make important decisions based on solid facts rather than their intuition. Predictive and prescriptive analytics can be used in cases where soil profiling [6, 7] is required and suggestions can be made to the farmer about which crop should be grown in a particular type of soil to maximize productivity. AI can also assist in estimating the timeline that farmers should follow to make use of the soil and the conditions perfectly in their favor.

Weather forecasts can also be very helpful to farmers to prepare better for an unforeseen conditions like a cyclone or any other natural disaster that may affect their productivity [8]. This will help them to reduce their losses which would have been huge without the help of AI and Satellite Imagery. In figure Fig. 9 we can see the use of weather forecasts to estimate the effect of a heatwave across The United States of America.

All of these features of crop yield prediction combine to give a holistic overview of the suitable time frames for crop production. There are also what-if scenarios, and alternative action plans to tackle any unforeseen problems. Businesses from the AgTech industry today are making use of neural network algorithms to predict crop yield. The backpropagation algorithms help in identifying the appropriate weight value of the yield to calculate the error derivative. Accuracy of crop yield estimation is significant for agronomic production reasons. So, predicting crop yield is essential for the food production ecosystem around the world. With better data in hand, it

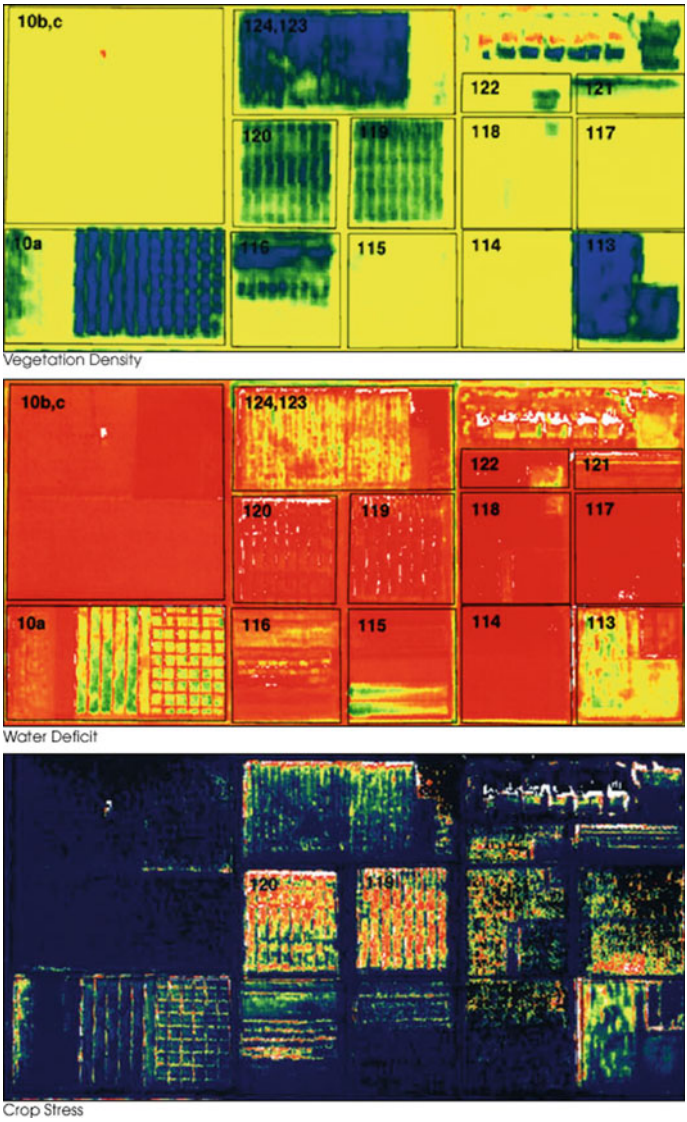


Fig. 7 Use of satellite imagery in precision agriculture

becomes possible to make informed decisions. Government agencies also find the crop yield prediction data useful as they can plan accordingly for National Food Security.

Fig. 8 Satellite imagery for monitoring of crops

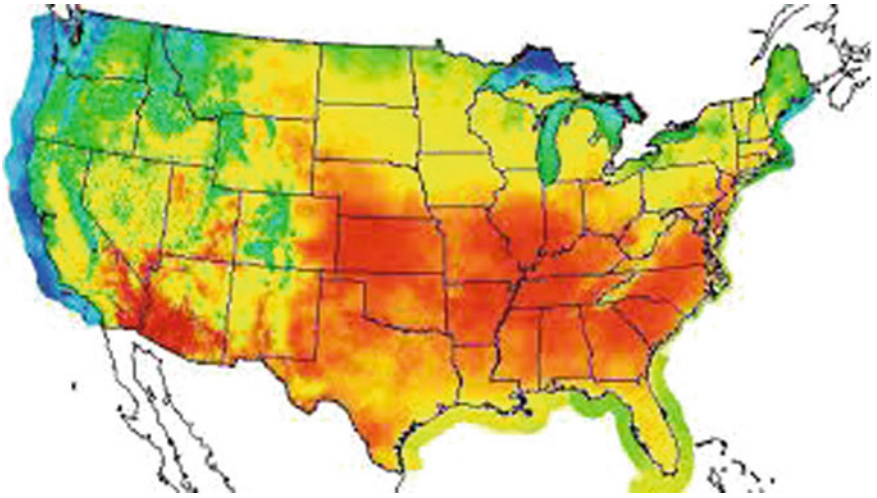


Fig. 9 Use of weather forecast in agriculture

3.2 Visibility and Control of Crop Yields

Low yield is one of the major problems worldwide that prevent the growth of agriculture. Especially in developing countries like India, where poor farming infrastructure and old agricultural practices are still prevalent have very low visibility and control on crop growth. The farm sizes are small with no proper use of technology. In these scenarios, AI and ML along with other IoT devices would be very beneficial. Its use can be seen in Fig. 10.

For instance, AI can be harnessed to suggest the amount of pesticide that is required and under precision farming [9] instruct the farmer on the areas of the farm that need special attention due to some unexpected conditions. Small farm sizes also mean that the resources like irrigation facilities, weed prevention policies, and financial support remain limited along with many other restrictions [10].

In addition to that, a small fluctuation in climatic conditions is enough to lead this already unstable industry toward a disaster. This may impact the yield heavily which directly affects cash flow. This makes it very difficult to allow future investments in this domain to improve productivity and mitigate future risk possibilities. There is a need for AgTech companies to focus on solutions that help predict yield based on factors like climate, use of pesticides, and many others. They should be able to recommend the amount of irrigation that is required for the farm keeping in mind the amount of rainfall and the climatic conditions predicted by the AI model.

The use of drones to regulate irrigation, supply of pesticides, and other forms of medicine to enhance the productivity of the farm is very common in many countries. This not only makes the farm more effective but also makes the applications of those chemicals more efficient. It reduces the wastage of chemicals and protects the land

Fig. 10 Use of AI for precision agriculture

Fig. 11 Use of drones and AI in precision agriculture

from overuse of chemicals which may affect its productivity. We can see their use in Fig. 11.

Thus, we can say that climate plays a very vital role and has a huge impact on agriculture as a whole. Changes in average temperature, rainfall levels, periodic heatwaves, carbon dioxide levels in the atmosphere, and ozone concentration at ground levels are a few of the climatic factors. Climate changes affect all countries unevenly. The countries with low altitudes are likely to face more issues with crop production. In Fig. 12 we can see the impact climate change and global warming has on rice farms in India.

There is a need for AgTech companies to predict the fluctuation in climatic conditions. They can harness historical data to establish a pattern between different seasons

Fig. 12 Impact of global warming on rice plantation

and provide insights like the amount of rainfall or the rise in temperature or humidity. There is also a possibility of predicting summer temperatures based on the winter data. If there is a fluctuation in summer temperatures, likely, winters will also have some change. These kinds of information are very beneficial to the farmer to make an informed decision about irrigation and other things that may help in increasing the productivity of his farm.

3.3 Soil Nutrient Depletion

Most nutrients in the soil are present in the top layer. This layer is subject to erosion over time which may lead to the nutrients getting depleted from the soil. This may lead to the solid becoming unfit for cultivation. It is a very serious problem faced by farmers all over the world. It affects the soil quality and is a big threat to the sustainability of the industry in the long term. In the image below, Fig. 13 shows how the erosion of the top layer can render a part of or sometimes the whole land unusable for the cultivation of any kind of crop.

AI and ML can play a vital role in tackling such an issue. Many novel algorithms are getting created that use state-of-the-art technology. They can help the farmers

Fig. 13 Erosion affecting productivity

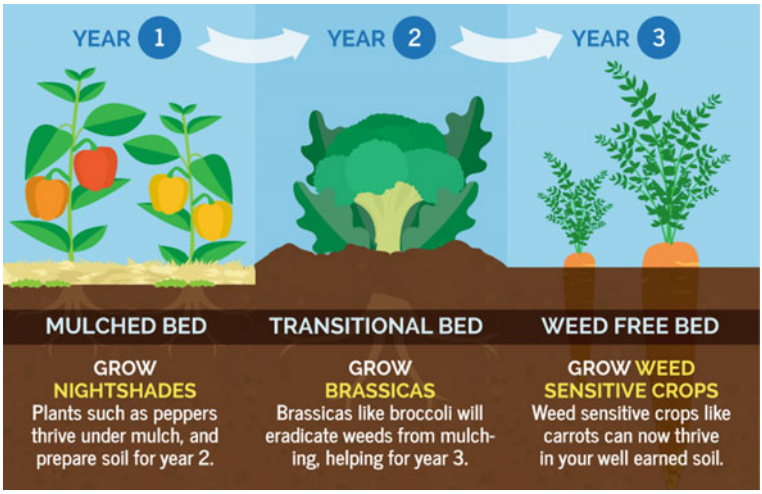


Fig. 14 Common technique of crop rotation

with modern techniques to prevent soil erosion. Soil sensors can be placed at various strategic locations throughout the farm, these can be used to detect nutrient levels. These sensors analyze the soil condition and generate timely alerts for preventive actions to be taken by the farmer. This is essential for monitoring soil and ultimately helps in improving crop yield. AI can also be harnessed to assist the farmers. It can recommend what kind of crop is best suited for a particular soil profile after thorough analysis. This is called prescriptive analytics which is very useful to farmers. It can substantiate the intuition of farmers with facts and data combined with various analytical tools.

Many crops aid in increasing the nutrient value of a particular land. This property of crops combined with advanced analytics, artificial intelligence, and fine-tuned recommendation systems can be used to support the farmer in choosing specific crop rotation techniques as shown in Fig. 14. This will be beneficial for the farm. It will be able to gain nutrients that can increase the productivity of the next harvest. It helps in building a sustainable farm with the help of technology.

3.4 Small Landholdings Due to Dense Population

It is a problem in developing countries like India because of its dense population and intensive cultivation. There are very small areas of land that are distributed between various people. This reduces the efficiency of the land to produce crops and also reduces the area of land that can be plowed or cultivated on. One of the reasons behind it is the inheritance laws. Farmers have to distribute their land equally among

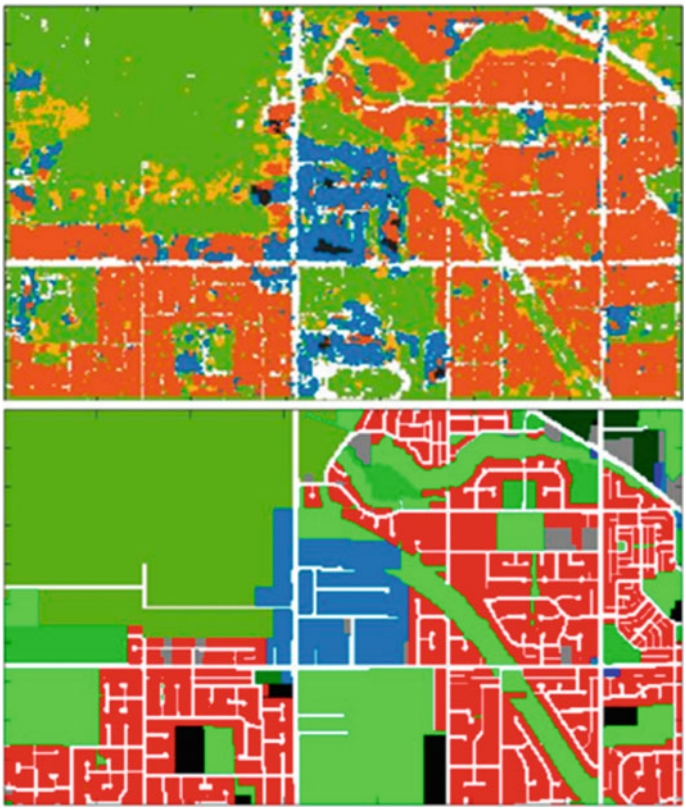


Fig. 15 Use of CNN for land use detection and land fragmentation

their sons. The distribution leads to fragmentation. This leads to a disparity in the level of fertility of different tracts.

The best solution to avoid this problem is to avoid fragmentation of the land. However, when that is not possible, the next best option is to have a system for zone detection and geo-tagging. Satellite imagery data can monitor soil conditions for identifying the best scenarios for crop production. It further helps in checking the natural environment of a farm for the targeted use of fertilizers.

In Fig. 15 we can see the use of satellite imagery and convolutional neural networks (CNN) to perform land use detection and perform land fragmentation accordingly.

3.5 Impact of Floods and Natural Disasters

Floods lead to the washing away of a lot of agricultural lands, leading to widespread losses both in production and infrastructure. Water Logging has its biggest impact on

Fig. 16 Impact of floods on farmlands

places where crops are planted. Crops do not survive under a heavy deluge of water. So, when production gets affected, it also leads to scarcity and famine. It ultimately leads to inflation in the prices of agricultural produce. This is not good for the people in the country as well as its economy takes a hit. Figure 16 shows the impact that floods can have on farmlands.

Weather sensors that predict uneven patterns of rainfall are essential to overcome this problem. They will send timely alerts that will help farmers propose mitigating measures and implement safeguarding policies to help them mitigate the risk. It can include the creation of soil bunds to protect the field from unexpected floods. They can build structures that manage the excess water efficiently as they are aware of the amount of rainfall through predictive analytics and weather forecasts.

Flood modeling and forecasting is a very important topic in the field of AI. In Fig. 17, we can see an example of GIS being used to map the flow or spread of a flood before it happens [11]. Climate change is responsible for the increase in the frequency and intensity of heavy rainfall. AI can serve as a remedy in the scenarios of floods and provide early and real-time warnings for these events to the government and the farmers. Technology is also used to simulate the risks in an area due to an impending flood [12]. Then the government can act on that information to save the people and the produce in that area.

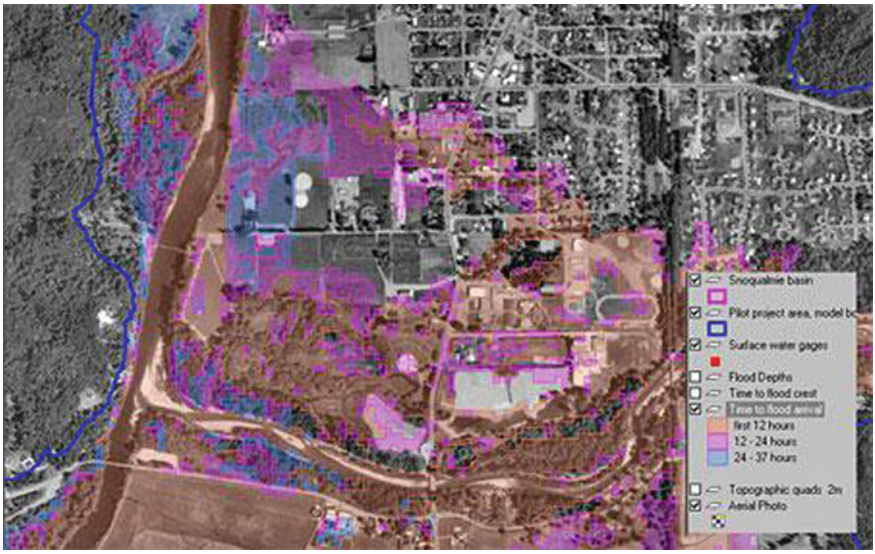


Fig. 17 Use of GIS to map flood before it happens

4 Tools and Techniques for Optimization in Agricultural Growth

There are many tools and techniques that are being used in the development of AgTech. These developments will help in the optimization of agricultural growth across the world and more importantly in developing nations that depend on agriculture for their economic growth. A few of them are mentioned in figure Fig. 18.

Geo-Spatial Analysis is used to add timing and location to traditional types of data and to build data visualizations. These visualizations can include maps, graphs, statistics, and cartograms that show historical changes and current shifts. This additional context allows for a more complete picture of events. It is used for data capture to understand anything from weather modeling to farmland segmentation. Geospatial big data analytics breaks data out of the endless rows and columns of a traditional spreadsheet and organizes it visually by time and space.

Geographical Information System or GIS is a computer system for capturing, storing, checking, and displaying data related to positions on Earth's surface. By relating seemingly unrelated data, GIS can help individuals and organizations better understand spatial patterns and relationships. It integrates common database operations, such as query and statistical analysis, with the ability to see how data relate in space and time. GIS along with Geo-Spatial Analysis can help in crop health monitoring, analysis, and yield protection to aid the farmers in making decisions to mitigate the losses and reduce the risk.

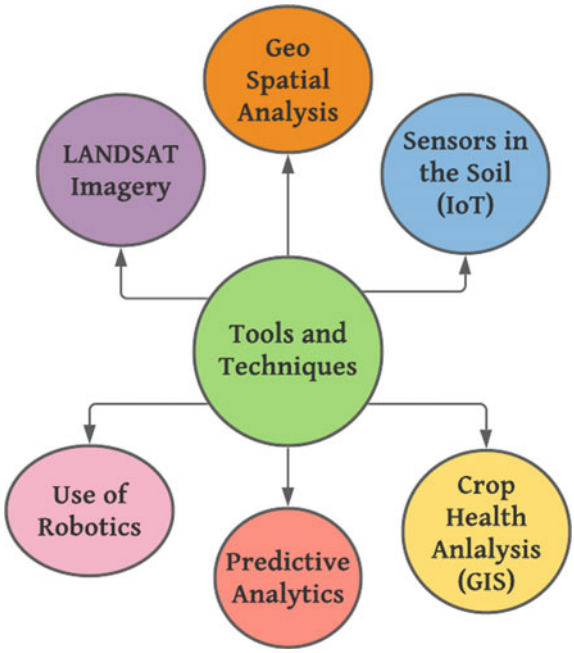


Fig. 18 Tools and techniques for optimization of agricultural growth

Sensors in the soil and other applications of IoT help in soil profiling. They can also be used to measure the moisture content and pH level in the soil. An application integrated with this can provide information on the various sections of the field, crop health and suggest the amount of manure, fertilizer, or pesticides that are required for different areas of the farm in real time.

This information along with that from drones, **advanced robotics**, and **predictive analytics** [13] can help in suggesting methods of crop rotation, irrigation management techniques, handling pest attacks, nutrition management, and much more.

The best source of this data is the **LANDSAT data** which has been providing genuine data from satellites since 1972. Whether it is about estimating crop production or monitoring water usage, LANDSAT imagery is of great importance. Besides, they are also beneficial in field-level management to identify different conditions and increase yield through zone mapping.

5 Conclusion

The main focus of this chapter is to give an overview of the growth of technology and its impact on agriculture. As discussed, in this chapter one can find the various difficulties in the agriculture industry for farmers like soil nutrient depletion, irregular

rainfall, the impact of natural disasters, and many more. In addition to that, they can find the tools and techniques that data scientists use in AgTech companies to aid the farmers and keep them informed.

The above work discusses the concepts of Geo-Spatial analysis, use of drone and satellite imagery to mine reliable information, use of LANDSAT data to do predictive and prescriptive analysis, and the use of AI and ML in the agriculture Industry to provide support to the farmers.

References

1. Kakani, V., Nguyen, V.H., Kumar, B.P., Kim, H., Pasupuleti, V.R.: A critical review on computer vision and artificial intelligence in food industry. J. Agric. Food Res. **2** (2020)
2. Shelestov, A., Kravchenko, A., Skakun, S., Voloshyn, S., Kussul, N.: Geospatial information system for agricultural monitoring. Cybern. Syst. Anal. **49** (2013). https://doi.org/10.1007/s10559-013-9492-5
3. Hoffmann, M., Butenko, Y., Traore, S.: Evaluation of satellite imagery to increase crop yield in irrigated agriculture. Agris On-Line Pap. Econ. Inform. **10**, 45–55 (2018). https://doi.org/10.7160/aol.2018.100304
4. Talaviya, T., Shah, D., Patel, N., Yagnik, H., Shah, M.: Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. Artif. Intell. Agric. **4**, 58 (2020)
5. Shafi, U., et al.: A multi-modal approach for crop health mapping using low altitude remote sensing, Internet of Things (IoT) and machine learning. IEEE Access **8**, 112708–112724 (2020). https://doi.org/10.1109/ACCESS.2020.3002948
6. Ayaz, M., Ammad-Uddin, Z., Sharif, A., Mansour, Aggoune, E.-H.M.: Internet-of-Things (IoT)-based smart agriculture: toward making the fields talk. In: IEEE Access, vol. 7, pp. 129551–129583 (2019). https://doi.org/10.1109/ACCESS.2019.2932609
7. Lepenioti, K., Bousdekis, A., Apostolou, D., Mentzas, G.: Prescriptive analytics: a survey of approaches and methods. In: BIS 2018 International Workshops, Berlin, Germany, July 18–20, 2018, Revised Papers (2019). https://doi.org/10.1007/978-3-030-04849-5_39
8. Frisvold, G., Murugesan, A.: Use of Weather information for agricultural decision making. Weather., Clim., Soc.. **5**, 55–69 (2013). https://doi.org/10.1175/WCAS-D-12-00022.1
9. Mokaya, V.: Future of precision agriculture in india using machine learning and artificial intelligence. Int. J. Comput. Sci. Eng. **7**, 1020–1023. https://doi.org/10.26438/ijcse/v7i2.10201023
10. Kumar, S.: Artificial intelligence in Indian irrigation. Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol. 215–219 (2019). https://doi.org/10.32628/CSEIT195536
11. El Morjani, Z.E.A., Ennasr, M., Elmouden, A., Idbraim, S., Bouaakaz, B., Saad, A.: Flood Hazard Mapping and Modeling Using GIS Applied to the Souss River Watershed. (2016). https://doi.org/10.1007/698_2016_69
12. Ullah, S., Farooq, M., Sarwar, T., Tareen, M., Wahid, M.: Flood modeling and simulations using hydrodynamic model and ASTER DEM—A case study of Kalpani River. Arab. J. Geosci. **9** (2016). https://doi.org/10.1007/s12517-016-2457-z
13. Kumar, V., L.M.: Predictive analytics: a review of trends and techniques. Int. J. Comput. Appl. **182** 31–37 (2018). https://doi.org/10.5120/ijca2018917434

Plenus Solutionem: A Complete Solution to Multi Chronic Diseases Detection and Classification

Aindril Kar, Harshit Anand, and Kalipada Tripathy

Abstract The present science and technology is so newfangled that it gives humans the freedom to talk about the sustainable development of resources not only on this planet but also on others. We are touching skies every new day. Advancement in technology has made it possible for people to live a peaceful life bringing down everything to the fingertips. But alas! this expansion isn't harnessed to its full. Despite these developments, we face a lot of challenges in the sector of health care. People have to stand in long queues even for the detection of their diseases via pathological tests, which is drawn-out and time-taking. Further, looking for a specialist is hectic. Moreover, the bills related to medical services are adequately adequate to gobble up all the money. Henceforth, the trained and deployed models under the realms of data science and machine learning are the way out. We have a number of data science prototypes on health care but the idea of this chapter is to provide better and more accurate predictions by creating a Generic model which is said to be trained on five different datasets of various types of diseases. The multi-chronic diseases that will be predicted by the model are—breast cancer, diabetes, chronic kidney disease, liver disorder, cardiovascular disease. The above diseases can be predicted accurately using the prototype and in less time with more reliability. This is to bring the understanding that machine learning isn't only about fitting and predicting values using some great bagging or boosting algorithms but about implementing a detailed machine learning pipeline. It is tried to bring out that great results are possible even using the linear or distance-based algorithms, if the data undergoes rigorous standardized and normalized preprocessing.

Keywords Data science · Machine learning · Breast cancer · Diabetes · Chronic kidney disease · Liver disorder · Cardiovascular disease · Logistic regression · KNN · Decision trees · XG boost · Normalization · Standardization

A. Kar · H. Anand (✉)
Highradius Technologies, Hyderabad, India
e-mail: harshitanand99@gmail.com

K. Tripathy
Rakuten Inc., Tokyo, Japan

S. S. Rautaray et al. (eds.), *Data Science in Societal Applications*, Studies in Big Data 114, https://doi.org/10.1007/978-981-19-5154-1_8

1 Introduction

A prototype that uses the principles of data science to foresee the disease of a given patient based on its symptoms is tried to be implemented. A generic model for five diseases is developed which can be further implemented as a tool. Datasets from different sources are trained to achieve the desired accuracy and efficiency. To predict the disease different regression techniques are performed on the dataset. The diseases are breast cancer, diabetes, chronic kidney disease, liver disorder, cardiovascular disease.

Breast cancer: As per facts, breast cancer could be known as the most abundant disorder for the women in the States. Breast cancer can happen in all kinds of people, yet it's undeniably more normal in ladies. While cancer growth impacts individuals, all things considered, races, identities, and genders, it doesn't generally influence them similarly. Contrasts in hereditary qualities, chemicals, ecological openings, and different elements can prompt contrasts in hazard among various gatherings of individuals. For most cancer, however, expanding age is the main danger factor. Research has shown the benefits of physical activities in women with breast cancer who have participated in exercise programs before and after cancer treatment. Women who participated in an exercise reported feeling better both mentally and physically. Non-cancerous breast tumors are abnormal growths, but they are not life-threatening.

Diabetes: Glucose in our body is the most important wellspring and source of energy which comes from the food we eat. When this glucose level rises excessively in our bodies, we get to suffer from diseases and disorders like the diabetes. As per the records, 463 million adults were diagnosed with Diabetes in 2019 and the numbers are estimated to rise to 700 million by 2045. Type 2 diabetic patients are on an increase day by day. More than 1.1 million children and adolescents are said to be living with Type 1 diabetes. As per IDF, Diabetes has caused around 4.2 million deaths as of now. Around USD 760B has been expended on health care for diabetes in 2019 which is 10% of the total spending on adults. Type 1, Type 2, gestational diabetes, impaired glucose tolerance, and impaired fasting glycemia are a few of the most common types of diabetes. With a person being prolonged diabetic, it could be harmful to the heart, kidneys, eyes, blood vessels, and nerves. There aren't many cures for the diabetes, few of them includes having a healthy diet and an active workout which could help bringing down the glucose and sugar level. Less use of Tobacco is also encouraged.

Chronic Kidney Disease: Chronic Kidney Disease (CKD) is a condition portrayed by a progressive loss of kidney work over the long run. In case kidney illness deteriorates, squanders can work to significant levels in the blood and cause us to feel debilitated. A bad kidney may lead to the accumulation of wastes in our body which would eventually lead to further diseases and disorders. The risks of inculcating issues in the heart and veins also increase over a period of time. But to the rescue, an early and regular check-up can be a boon in stopping our kidney from further deteriorating. At the point when kidney illness advances, it might ultimately prompt kidney

disappointment, which requires dialysis or a kidney relocate to keep up with life. Frequently, however, chronic kidney disease has no fix, therapy ordinarily comprises measures to assist with controlling signs and side effects, diminish difficulties, and slow the movement of the illness.

Liver Disorder: The liver does a ton of things that keeps us sound. It transforms supplements into synthetics our body needs. It sifts through harm. It helps transform food into energy. Cirrhosis is yet another complex version of Liver disorder which is similar to the liver fibrosis and is joined by the hepatic vasculature. Reasons for cirrhosis can generally be distinguished by the patient's set of experiences joined with the serological and histological examination. Alcoholic liver infection and hepatitis C and B are the most widely recognized reasons for cirrhosis.

Heart Disease: The CDC also known as the Centers for Disease Control says, heart illness is the main source of death in the United States, Canada, United Kingdom, and Australia. One in every four passings happens because of heart illness. Inherent heart illness is an overall term for certain distortions of the heart that have been available since birth. Arrhythmia is a sporadic heartbeat. The coronary veins supply the heart muscle with supplements and oxygen by flowing blood. Few of the common symptoms which could be very helpful in characterizing the heart disease are heart palpitations, breathlessness, and pain in the chest. The chest pain which occurs in most people suffering from heart disease is because of insufficient and irregular supply of oxygen in all parts of the heart. Heart diseases generally aren't hereditary with an exception of hyper-trophic cardiomyopathy. Hypertension is assessed to represent roughly 13% of coronary illness passings, while tobacco represents 9%, diabetes 6%, absence of activity 6%, and heftiness 5%. Cardiovascular infections are the main source of death in every aspect of the world aside from Africa. Together, cardiovascular sicknesses have brought about 17.9 million passings in 2015.

We performed various data regression techniques on the given dataset to achieve better accuracy and efficiency. Logistic regression: Calculated regression is a machine learning calculation that is utilized for classification issues, it is a predictive investigation calculation and in light of the idea of probability. Logistic regression is named on the logistic function which it uses to display a binary dependent variable. The comparing probability of the value named '1' can shift somewhere in the range of 0 and 1, consequently the marking; the capacity that converts log-chances to probability is the logistic capacity.

KNN: K-Nearest Neighbors (KNN) is as far the comparatively easiest calculation utilized in machine learning for regression and classification statements. KNN calculations use the data and arrange new data points dependent on closeness measures. KNN works by tracking down the distances between a query and every one of the models in the data, choosing the predefined number of models (K) nearest to the question.

Decision Tree: The decision tree algorithm is a type of supervised learning algorithm and can be used in both classification and regression problems. The decision tree utilizes the tree portrayal to solve the problem statement in which each base node

compares to a class name and features are addressed on the interior nodes of the tree. Overfitting is the peculiarity where the learning framework firmly fits the given preparation date such a lot that it would be mistaken in foreseeing the results of the unlearned data. In decision trees, overfitting happens when the tree is intended to perfectly fit all examples in the preparation data collection.

XG Boost: XG Boost is an algorithm that has as of late been overwhelming applied AI for structured or tabular data. XG Boost is an execution of decision trees that are gradiently boosted and intended for speed and execution. It is an adaptable and exact execution of gradient-boosting prototypes that have demonstrated to stretch the boundaries of figuring power for boosted tree algorithms.

2 Literature Survey

In the last couple of years, many research works are carried in the space of health care forecasting utilizing different ideas and calculations of data science, AI, and machine learning. An escalated study on the exploration of such papers which centers around the comparative area is done under this section.

Wei et al. [1] exclaimed a nomogram for the heart disease prediction for the patients having rheumatoid arthritis by finding the F1-score of machine learning models such as logistic regression, decision tree, K-nearest-neighbors, XG gradient-boosting, random forest, and SVM with increasingly stable execution. Ye et al. [2] proposed a danger alleviation tool utilizing machine learning-based calculations like XG boost and noticed the data accumulated from the electronic well-being records containing subtleties of more seasoned patients and concluded that the device achieved a further developed discriminative limit which can speedily be sent in the medical care framework to give programmed early caution to more established patients with expanded fall hazard. Anand et al. [3] have tried to bring out the importance of technology in the field of medical and health care by developing a prototype that can predict and foresee the type and occurrence of heart disease based on parameters like age, sex, resting blood sugar, etc. The research work seems to be confiscated with the theoretical findings and seems to have a lot of scope for the deployed model. However, the performance could have been more efficient with the current bagging and boosting algorithms.

Nusinovici et al. [4] have executed five diverse AI calculations, for example, SVM, gradient boosting, KNN, single-hidden-layer and random forest which were contrasted with logistic regression dependent on execution for the forecast of cardiovascular illnesses, persistent kidney sickness, diabetes, and hypertension with basic clinical boundaries. Further, they observed that the best performing models were logistic regression and neural networks. Thapa et al. [5] came up with two computational and logical methods DL-MaloSite and RF-MaloSite, based on machine learning algorithms for the predictions in proteins based on their amino acid sequences in the primary category and found that the DL-MaloSite algorithm

performs comparatively better than the RF-Malosite algorithm upon having an MCC score of 0.51 and 0.49. But considering the size of the dataset, the model had a high scope of improvement in the prediction system. Also, a multi-windows input could have been used for the up-gradation. Maini et al. [6] had worked on an approach to remove the reiterative and immaterial features and to distinguish the significant attributes from the Cleveland heart disease dataset by using few feature selection algorithms such as the Filter Feature Selection algorithm (MRMR, Relief), Wrapper (Genetic algorithms) and embedded feature selection algorithms. Further, the prediction model was built by applying algorithms of machine learning such as logistic regression, k-nearest neighbors, random forest, naive byes, and support vector machine. But the performance of the system could have been progressed by actualizing other feature selection techniques like forward feature selection and backward feature elimination as well, to further increase the accuracy by reducing the computational power and time.

Kumar et al. [7] brought out the detection of Coronary Artery Disease (CAD) using Spectral and Spectro Temporal Techniques. The different features of heart sounds are extracted using an electronic 3M Littmann 3200 stethoscope. Stockwell transformation is performed followed by bandwise spectral kurtosis. Finally, the classifying bandwidth is found using the minimum distance classifier. Porkodi and Karuppusamy [8] came up with a system implementing Gabor Filtration along with Random Forest Classification (RGRFC). The adopted techniques have been tested for their specificity, efficiency, precision, and sensitivity on COPD Datasets. However, Gabor Filter comes with few demerits of being unstarred in nature, which means it requires to be done from different angles to gain a superimposed image for improving accuracy. Also, the attribute weights generated by Random Forest on the data seem not to be credible. Henceforth, the accuracy accomplished has more scope for improvement.

Ming et al. [9] have tried increasing the prediction score for the machine learning models which are used in health care and medicals and have low discriminatory accuracy to Breast Cancer Risk Assessment Tool (BCRAT) and Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) models. A significant rise in predictive accuracy has been observed using machine learning adaptive boosting and machine learning Markov chain Monte Carlo generalized linear mixed models. They further stress the fact that higher accuracies are very important in the fields of medicine. Li et al. [10] have looked through the PubMed (counting MEDLINE), Embase, and Web of Science Core datasets for anticipating the 5-year endurance pace of Breasts Diseases. They further shout that contrasted with conventional statistical strategies, the presentation of current ML models doesn't really show any improvement, and this space of exploration actually faces limits identified with an absence of data preprocessing steps, the very high contrasts of test feature selection, and issues identified with validation. Also, improvement of the presentation of the proposed model is likewise required later on, which requires more standardization, normalization, and likewise validation testing.

Ekanayake and Herath [11] have proposed a work process to foresee the status of CKD, consolidating data preprocessing, a missing value handling technique with a

collaborative feature selection. Out of the eleven AI strategies they have thought of, the extra tree classifier and RF classifier are displayed to bring about the most noteworthy precision and negligible bias to attributes. The exploration likewise thinks about the pragmatic parts of data collection and features the significance of consolidating domain knowledge when utilizing AI for CKD status expectations. Reshma et al. [12] propose the use of Support Vector Machine (SVM) classifier and Ant Colony Optimization (ACO) technique for a chronic kidney disorder to find the occurrence of the disease using a minimal set of features while maintaining a higher accuracy of around 96%.

Mujumdar and Vaidehi [13] have proposed a diabetes forecast model for better arrangement of diabetes which incorporates not many outside factors liable for diabetes alongside ordinary variables like Glucose, BMI, Age, Insulin, and so on Order exactness is helped with the new dataset contrasted with the current dataset as of January 2019. Further a forced pipeline model for diabetes expectation is planned toward working on the exactness of accuracy. Soni and Varma [14] have proposed an AI model utilizing the characterization and ensembling methods like support vector machine, decision trees, gradient boosting, KNN, random forest, and logistic regression. Out of the various models they tracked down random forest classifier to think of the best precision. Furthermore, this exploration work appeared to have a ton to go as far as data preprocessing, feature engineering, and feature selection methods for better outcomes. Rahman et al. [15] have worked on evaluating the performance of different machine learning algorithms to reduce the high cost of chronic liver disease diagnosis by prediction using algorithms like KNN, SVM, decision trees, random forest, and Naive Bayes. Accuracy, precision, recall, f-1 score, and specificity are the evaluation metrics for different classification techniques. The analysis result showed the logistic regression to achieve the highest accuracy. Notwithstanding, the review is predominantly centered around the utilization of clinical information just for liver infection expectations.

3 The Datasets Explained

Breast Cancer Wisconsin Dataset: Highlights are figured from a digitized picture of a fine needle suction of a Breast mass. They depict the attributes of the cell cores present in the picture. Ten genuine esteemed highlights are processed for every cell core in this dataset; compactness, concavity, perimeter, concave points, fractal dimension, radius, smoothness, area, symmetry, and texture. *Mammographic Mass Dataset:* Separation of harmless and threatening mammographic masses dependent on BI-RADS ascribes and the patient's age are available in this dataset. Mammography is the best technique for breasts disease screening accessible today. Nonetheless, the low sure prescient worth of breasts biopsy coming about because of mammogram understanding prompts roughly 70% superfluous biopsies. The dataset comprised 961 cases with 6 credits altogether.

Pima Indian Diabetes Dataset: It was taken in the year 1990 with a total of 9 attributes with 768 instances. The dataset was generated by the National Institute of Diabetes and Digestive and Kidney Diseases. *Chronic Kidney Disease Dataset:* It had 400 instances in total and with 25 attributes like age, blood pressure, specific gravity, albumin, sugar, RBC, pus cell, pus cell clumps, bacteria, blood urea, sodium, potassium, hemoglobin, etc. *ILPD (Indian Liver Patient Dataset):* There are around 400 records of liver patients and 200 non-liver patients. The dataset was collected from the northeast of Andhra Pradesh, India. The informational index was gathered from the upper east of Andhra Pradesh, India. The selector is a class mark used to partition into gatherings. This dataset contains 441 male patient records and 142 female patient records. *Heart Disease Dataset:* This dataset is a combination of the heart disease statistics for the region of Cleveland, Hungary, Switzerland, and the VA Long Beach. It contains 303 instances with a total of 14 important attributes. *Public Health Heart Disease Dataset:* This is a Kaggle dataset and is taken from 1988 and contains 13 attributes. There are 1026 instances for the following with 14 attributes. *Cardiovascular Disease Dataset:* The dataset consists of 70,000 records of patients.

4 The Core Machine Learning Pipeline

The basic machine learning pipeline which is tried to be implemented in this research work is tried to be portrayed diagrammatically in Fig. 1.

Data Preprocessing and Statistical Analysis: This starts with the first step of identifying the dependent and the independent variables in the dataset according to the use case. It's a classification type with the labels as occurrence of a disease or non-occurrence of a disease. This is followed by finding all the NaN/Null values in the dataset corresponding to each column and replacing them with suitable values like the mean of the column if no outlier is present, otherwise replaced by the median. If a column contains null values more than a threshold percentage it is dropped. Duplicate rows if present is dropped. Further, handling of string and categorical data in the dataset is done. Since it becomes difficult for the machine learning algorithms to work directly with string data, therefore all the columns containing string data are the first label encoded, and then after analyzing the data if these columns are found to be nominal categorical (independent of each other) variables then these are further Onehot encoded, if they share an ordered relation then no such encoding is done to retain the relation. A detailed analysis of each column is done, i.e., the unique values of each column are found to check for the constant or quasi constant features. A statistical description of each attribute is done which helped to check the consistency of the data, understand the nature of the data and find out the anomalies. Then, the distribution of the data is being checked, if it followed a normal distribution or a bell-shaped curve is obtained which pertains that the data is said to be skewed. Various transformations like logarithmic, exponential, inverse, power, and Box Cox transformations are performed on the data to obtain a normal distribution and proper

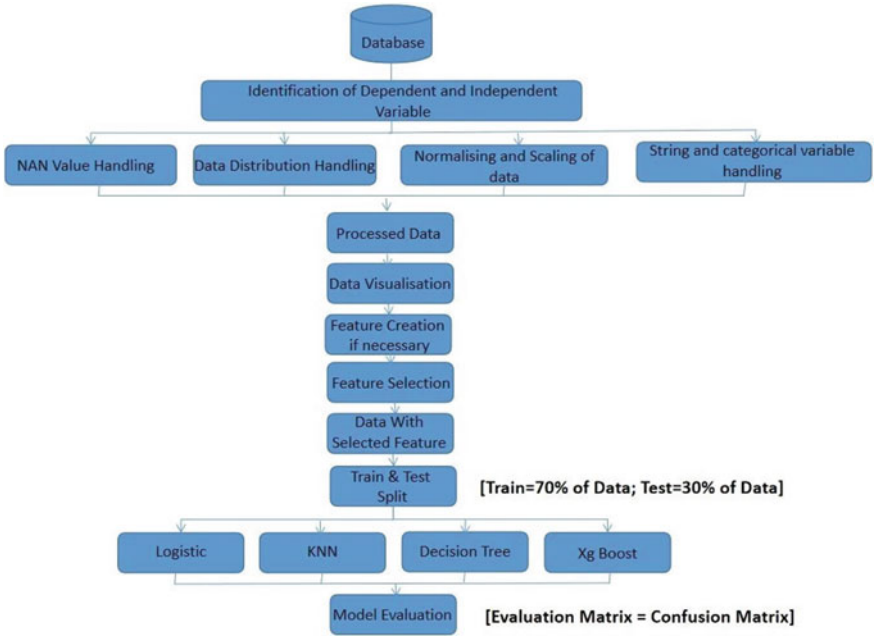


Fig. 1 Machine learning pipeline

handling of outliers. Scaling of the data is done such that no independent variable is said to be dominant over the others in producing the output variable due to its high numeric value range, for this purpose, the standard scaler was used, which means if x is an independent variable then x is scaled using the function:

$$X = (X - \text{Mean}(X))/\text{Standard Deviation}(X)$$

Data Visualization and Exploratory Data Analysis: A Pairplot is used to visualize the relationship shared by all the features among themselves as in Fig. 2. A correlation heat map in terms of a correlation matrix which is a table showing correlation coefficients between sets of variables is plotted in Fig. 3. The diagonal of the table is always a set of ones because the correlation between a variable and itself is always 1. This plot is used to find the correlation between the dependent and independent variables thereby helping to select the independent features which are highly correlated with the dependent variable.

Feature Engineering: New features which are found to possess a high correlation with the dependent variable such that a unit change in these features resulted in a huge change in the output variable are created. These features are created by a combination of various independent features. For example, in the Pima Indian Diabetes dataset, a feature 'Score' was created using the below expression:

Fig. 2 Pairplot explaining relationship between the features

$$\text{df['score'] } = \text{ (df['Glucose'] } * \text{ df['BMI']) / df['BloodPressure']}$$

This denotes the variation of glucose level with a unit change in blood pressure corresponding to a certain BMI such that the co-relation of this feature is found to be 0.41 with the output variable.

Feature Selection: The filter method is used in some datasets where a correlation heat map is drawn to find the correlation of all the independent features with dependent variables such that only the features with high correlation (greater than a threshold) with the output are selected for the modeling phase. The embedded method in which the features are selected on the basis of their feature importance is calculated by fitting the data to the tree-based algorithms having feature importance attribute is also implemented. Forward feature selection and backward elimination are also attempted which is an iterative process in which only one feature (highly correlated) is fit to the model and its performance is noticed, followed by an iteration in which another

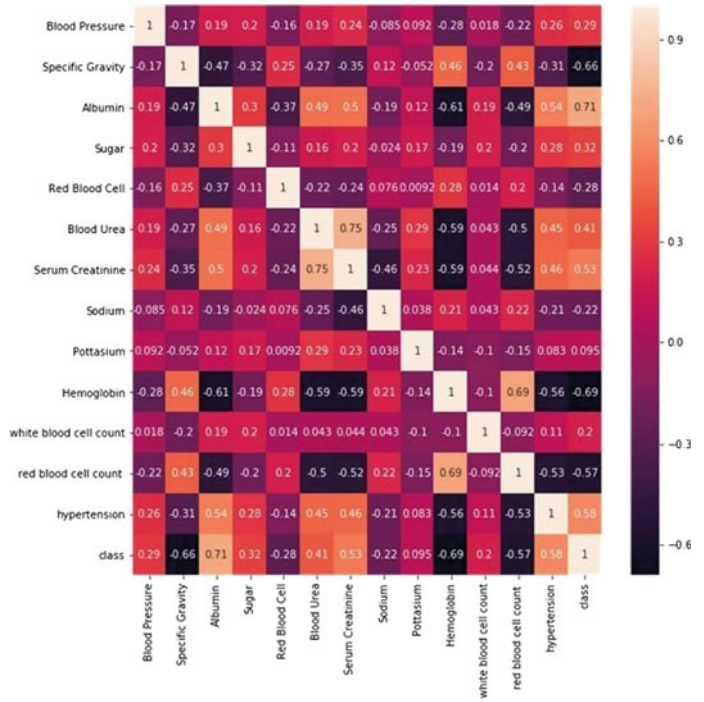


Fig. 3 Heatmap showing correlation among the features of chronic kidney disease dataset

feature is added or removed from the forming set of features and performance is measured. This process is continued till all the features are fitted. If on adding a feature the performance decreases, it is discarded. This eventually gives an optimal set of features giving the best performance. The above feature selection methods are implemented to bring an optimal set of features and their performance is measured and finally, a set of features is chosen.

Data Modeling: The processed data with selected features are then split into a train set and a test set using the Random Seed which is a random splitting keeping a check on the class count such that the class is always said to be balanced in the train set. The dataset was split in the ratio of 70% (Train) and 30% (Test). To perform a generic prediction of diseases, various algorithms are used. The algorithms are as follows:

Logistic regression: Logistic regression is a machine learning algorithm that is based on the probability of the predictive analysis. It is used to classify the problems. Generally this output of the models depends on the number of inputs and it's not a classifier. "Logistic regression", the name is coined after the function n which it operates, the logistic function. Logistic regression uses an equation as the representation, very much like linear regression. Below is an example logistic regression equation:

$$y = e^{(b0+b1*x)}/(1 + e^{(b0+b1*x)})$$

KNN: K-Nearest Neighbors (KNN) is a linear-distance-based supervised-machine learning algorithm used for regression and classification problems. The standard KNN algorithm works on the principle of euclidean distance such that for a given input data point it tries to estimate n nearest data points (similar data points) by estimating the euclidean distance of the given data point w.r.t. to the entire train data points and the output of the n nearest data points are then aggregated (mode/mean depending upon the use case) to foresee the result of that input test-data point. This can be thought of as the training set for the algorithm, though no explicit training step is required.

Decision Tree: A decision tree is an algorithm that arrives at a resultant, assigning some weights to the features involved and then characterizing and classifying them. It has a tree-like structure for decision-making. It's an application of conditional control statements. This algorithm can be used for solving regression and classification problems. This tree is basically used to predict the class by learning some simple decision rules based on the training data.

XG Boost: XG Boost has become a widely used and popular tool among Data Scientists. It is used for a scalable and accurate implementation of gradient-boosting machines. It is used for its high model performance and computational speed. XGBoost is an enhanced and circulated gradient-boosting library intended to be exceptionally productive, adaptable, and compact. It carries out AI and ML calculations under the Gradient Boosting structure. XGBoost gives a sequential boosting (otherwise called GBDT, GBM) that takes care of numerous machine learning issues in a quick and precise manner. The similar code runs on a significantly dispersed environment (Hadoop, SGE, MPI) and can tackle issues past billions of models (Tables 1, 2, and 3).

Evaluation Metrics:

	1(predicted)	0(predicted)	
1(actual)	TP	FN	Recall for class 1: TP/(TP + FN)
0(actual)	FP	TN	Recall for class 0: TN/(TN + FP)
	Precision for class 1:	Precision for class 0:	
	= TP/(TP + FP)	= TN/(TN + FN)	

5 Conclusion

The paper overviews the use of machine learning in the field of disease prediction. Different classification models of KNN, logistic regression, XG Boost, and decision tree are performed on disease datasets like breast cancer, diabetes, chronic kidney

Table 1 Comparative analysis of accuracy

Technique→ Disease↓	Logistic regression	KNN	Decision tree	XG boost
Breast cancer Wisconsin dataset	Train: 98.356% Test: 98.601%	Train: 98.122% Test: 95.804%	Train: 96.244% Test: 95.104%	Train: 99.530% Test: 98.601%
Mammographic mass dataset	Train: 82.465 Test: 84.155	Train: 82.986 Test: 83.116	Train: 83.333 Test: 82.857	Train: 84.201 Test: 83.636
Pima diabetes dataset	Train: 75.868% Test: 79.687%	Train: 78.298% Test: 78.125%	Train: 81.067% Test: 78.125%	Train: 82.465% Test: 81.250%
Chronic kidney disease dataset	Train: 98.92% Test: 97.50%	Train: 97.14% Test: 96.66%	Train: 100% Test: 98.33%	Train: 100% Test: 100%
Liver patient dataset	Train: 74.754% Test: 71.428%	Train: 75.735% Test: 72.571%	Train: 73.774% Test: 72.571%	Train: 74.754% Test: 72.771%
Public heart disease dataset	Train: 83.821% Test: 85.064%	Train: 85.216% Test: 82.792%	Train: 90.376% Test: 87.987%	Train: 99.302% Test: 98.051%
Cardiovascular disease dataset	Train: 100% Test: 100%	Train: 99.93% Test: 99.89%	Train: 100% Test: 100%	Train: 100% Test: 100%

Table 2 Confusion metrics for the datasets. Here label = [1, 0] signifies the confusion matrix above this table

Technique→ Disease↓	Logistic regression	KNN	Decision tree	XG boost
Breast cancer Wisconsin dataset [0, 1]	[97, 0 2, 44]	[94, 3 3, 43]	[91, 6 1, 45]	[96, 1 1, 45]
Mammographic mass dataset [1, 0]	[157, 23 38, 167]	[154, 26 39, 166]	[157, 23 38, 167]	[158, 22 41, 164]
Pima diabetes dataset [1, 0]	[41, 26 13, 112]	[42, 25 17, 108]	[49, 18 24, 101]	[43, 24 12, 113]
Chronic kidney disease dataset [0, 1]	[47, 1 2, 70]	[47, 1 3, 69]	[48, 0 2, 70]	[48, 0 0, 72]
Liver patient dataset [1, 2]	[111, 17 33, 14]	[105, 23 25, 22]	[111, 17 31, 16]	[117, 11 38, 9]
Public heart disease dataset [1, 0]	[140, 14 32, 122]	[128, 26 27, 127]	[147, 7 30, 124]	[152, 2 4, 150]
Cardiovascular disease dataset [0, 1]	[10488, 0 0, 10215]	[10473, 15 7, 10208]	[10488, 0 0, 10215]	[10488, 0 0, 10215]

disease, liver disorder, and chronic heart disease. Many previous representational research papers could be compared to our research work based on the classification models, but the goal of our work is to achieve higher accuracy and present an optimal, efficient, and practically possible solution to society. For the cardiovascular disease dataset, the highest accuracy achieved we could achieve was 100% using the XG Boost learning algorithm as compared to the rest of the algorithms. In order to ensure that the scores populated aren't having any bias and the model isn't overfitting,

Table 3 Precision and recall for each dataset

Technique→ Disease↓	Logistic regression	KNN	Decision tree	XG boost
Breast cancer Wisconsin dataset	Precision class *: 0.97 class #: 1	Precision class *: 0.96 class #: 0.93	Precision class *: 0.98 class #: 0.88	Precision class *: 0.98 class #: 0.97
	Recall class *: 1 class #: 0.95	Recall class *: 0.96 class #: 0.93	Recall class *: 0.93 class #: 0.97	Recall class *: 0.98 class #: 0.97
Mammographic mass dataset	Precision class #: 0.80 class *: 0.87	Precision class #: 0.81 class *: 0.86	Precision class #: 0.80 class *: 0.87	Precision class #: 0.79 class *: 0.88
	Recall class #: 0.87 class *: 0.81	Recall class #: 0.85 class *: 0.80	Recall class #: 0.87 class *: 0.81	Recall class #: 0.87 class *: 0.80
Pima diabetes dataset	Precision class #: 0.75 class *: 0.81	Precision class #: 0.71 class *: 0.81	Precision class #: 0.67 class *: 0.84	Precision class #: 0.78 class *: 0.82
	Recall class #: 0.61 class *: 0.89	Recall class #: 0.62 class *: 0.86	Recall class #: 0.73 class *: 0.80	Recall class #: 0.64 class *: 0.90
Chronic kidney disease dataset	Precision class *: 0.95 class #: 0.98	Precision class *: 0.94 class #: 0.98	Precision class *: 0.96 class #: 1	Precision class *: 1 class #: 1
	Recall class *: 0.97 class #: 0.97	Recall class *: 0.97 class #: 0.95	Recall class *: 1 class #: 0.97	Recall class *: 1 class #: 1
Liver patient dataset	Precision class *: 0.77 class #: 0.45	Precision class *: 0.80 class #: 0.48	Precision class *: 0.75 class #: 0.48	Precision class *: 0.75 class #: 0.45
	Recall class *: 0.86 class #: 0.29	Recall class *: 0.87 class #: 0.46	Recall class *: 0.86 class #: 0.30	Recall class *: 0.91 class #: 0.19
Public heart disease dataset	Precision class #: 0.81 class *: 0.88	Precision class *: 0.82 class #: 0.83	Precision class *: 0.83 class #: 0.94	Precision class #: 0.97 class *: 0.98
	Recall class #: 0.90 class *: 0.79	Recall class #: 0.83 class *: 0.82	Recall class *: 0.95 class #: 0.80	Recall class #: 0.98 class *: 0.97
Cardiovascular disease dataset	Precision class *: 1 class #: 1	Precision class *: 0.99 class #: 0.99	Precision class *: 1 class #: 1	Precision class *: 1 class #: 1
	Recall class *: 1 class #: 1	Recall class *: 0.99 class #: 0.99	Recall class *: 1 class #: 1	Recall class *: 1 class #: 1

a sequential Hyperparameter Tuning is performed using different samples of the population. The Breast Cancer dataset could give the highest accuracy on the XG Boost learning algorithm at 98.60% which in comparison to previous research works, was 96.05%. The maximum accuracy obtained for the public heart disease dataset was 98.05% using XG Boost as compared to the rest of the models, the previous models on the heart diseases dataset we considered, the maximum achievable accuracy was 87%. For the chronic kidney disease dataset, the highest accuracy obtained was 100% using the XG Boost algorithm where the maximum around other previous works was around 96.75%. The Pima Indian Diabetes dataset could give the maximum accuracy of 81.25%. The Indian Liver Patient dataset had the maximum accuracy obtained at 72.77% using the XG Boost algorithm and according to the previous research papers, the maximum accuracy which could be obtained was 71.86%. The model can be used to predict the above diseases accurately and in less time with more reliability.

The doctors and the current medical resources are undoubtedly the first Goto available through all thick and thins. But they are blessed with some sure cons as discussed in the chapter above. Here, science and technology under the realms of machine learning and data science come to our rescue. According to stats, even the maximum skilled medical doctors can predict the incidence of cardiovascular diseases with a precision of 67% only. On the contrary, Machine learning algorithms can provide outcomes with accuracy ranging within the 90 s. Machine learning can go about as a benefit over the impediments of human dynamics and carelessness toward admittance to mind. These modern-day algorithms can be a method of disposing of long medical bills that too in lesser time than remaining in long queues before medical clinic counters. They can give results at whatever point and at any place required right at the mark of care. Just by asking for a couple of manifestations, symptoms, and parameters on a platform, and letting a model detect and classify the occurrence of disease can save a ton of time and in particular will be in the financial plan of commoners. Patients need the physical presence of the doctors to diagnose any disease, our objective is to create a generic model for disease prediction and it will be a service to the doctors, eventually increasing the efficiency in the field of health care. The rapid emergence and evolution of science and technology have enabled us to do away with the cons of the traditional methods of detecting diseases and disorders.

References

1. Wei, T., Yang, B., Liu, H., Xin, F., Fu, L.: Development and validation of a nomogram to predict coronary heart disease in patients with rheumatoid arthritis in northern China. Aging (Albany NY) **12**(4), 3190 (2020)
2. Ye, C., Li, J., Hao, S., Liu, M., Jin, H., Le, Z., Xia, M., Jin, B., Zhu, C., Alfreds, S.T., Stearns, F.: Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm. Int. J. Med. Inf. **137**, 104105 (2020)
3. Anand, H., Anand, A., Das, I., Rautaray, S.S., Pandey, M.: Hridaya Kalp: a prototype for second generation chronic heart disease detection and classification. In: Gupta, D., Khanna, A.,

Bhattacharyya, S., Hassanien, A.E., Anand, S., Jaiswal, A. (eds.) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol. 1166. Springer, Singapore
4. Nusinovici, S., Tham, Y.C., Yan, M.Y.C., Ting, D.S.W., Li, J., Sabanayagam, C., Wong, T.Y., Cheng, C.Y.: Logistic regression was as good as machine learning for predicting major chronic diseases. J. Clin. Epidemiol. (2020)
5. Thapa, N., Hiroto, S., Roy, K., Newman, R.H., Dukka, K.C.: RF-MaloSite and DL-Malosite: methods based on random forest and deep learning to identify malonylation sites. Comput. Struct. Biotechnol. J. (2020)
6. Maini, E., Venkateswarlu, B., Gupta, A.: Role of feature selection in building high performance heart disease prediction systems. ADBU J. Eng. Technol. **8**(2) (2019)
7. Kumar, G.G., Khan, S., Basha, M.: Preliminary diagnosis of coronary artery disease from human heart sounds: a signal processing prospective. Int. J. Adv. Trends Comput. Sci. Eng. (IJATCSE) **8**(3) (2019)
8. Porkodi, V., Karuppusamy, S.: Preliminary diagnosis of coronary artery disease from human heart sounds: a signal processing prospective. Int. J. Adv. Trends Comput. Sci. Eng. (IJATCSE) **8**(5) (2019)
9. Ming, C., Viassolo, V., Probst-Hensch, N., et al.: Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Res. **21**, 75 (2019)
10. Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., et al.: Predicting breast cancer 5-year survival using machine learning: a systematic review. PLoS ONE **16**(4), e0250370 (2021)
11. Ekanayake, I.U., Herath, D.: Chronic kidney disease prediction using machine learning methods. In: 2020 Moratuwa Engineering Research Conference (MERCon) (2020)
12. Reshma, S., Shaji, S., Ajina, S.R., Vishnu Priya, S.R., Janisha, A.: Chronic kidney disease prediction using machine learning. Int. J. Eng. Res. Technol. (IJERT) **9**(7) (2020)
13. Mujumdar, A., Vaidehi, V.: Diabetes prediction using machine learning algorithms. Procedia Comput. Sci. (2019)
14. Soni, M., Varma, S.: Diabetes prediction using machine learning techniques. Int. J. Eng. Res. Technol. (IJERT) **9**(9) (2020)
15. Rahman, A.K.M., Shamrat, F.M., Tasnim, Z., Roy, J., Hossain, S.: A comparative study on liver disease prediction using supervised machine learning algorithms (2019)

Kochen Helfer: An AI-Based Cooking Assistant

Indrashis Das, Anoushka Mishra, and Shushant Ghosh

Abstract The evolution of artificial ıntelligence has made the lives of humans effortless. It is difficult to imagine our lives without smartphones, video games, or smart home devices. Innovative kitchens are an integral part of smart home devices. The evolution of several kitchen gadgets has made cooking possible for even naive people. However, it becomes difficult for people to maintain the same quality and food standards since sometimes they don't have anyone to guide them. Kochen Helfer, an AI-based cooking assistant, is one such product that will be helpful to reach the quality of food desired through proper ımage processing, classification and recommendation. Kochen Helfer is a proposed application of Deep Learning that can be used by novices in the domain of cooking.

Keywords Artificial ıntelligence · Smart kitchen · Image processing · Recommendation system · Classification

1 Introduction

1.1 Intuition Behind the Proposed Product

Many of us have failed to follow a recipe properly and have ended up with an empty stomach many times. Even though the recipes are extremely detailed, we have ended up with maybe a burned, bland or uncooked meal. It makes us wonder, what if the recipes could talk to us and tell us that, "it's too late to save that curry, it's charred, start over".

The proposed product, Kochen Helfer, intends to replace the tedious recipes with a life-like cooking assistant. An assistant that can not only read out the recipes but also give real-time feedback on the state of the food being cooked. The primary

I. Das (✉) · S. Ghosh
HighRadius Technologies, Bhubaneswar, India
e-mail: indrashisdas98@gmail.com

A. Mishra
Dell Technologies, Bangalore, India

S. S. Rautaray et al. (eds.), *Data Science in Societal Applications*, Studies in Big Data 114, https://doi.org/10.1007/978-981-19-5154-1_9

intention behind this proposed product is to provide a human touch and inculcate a human-like response.

1.2 Answering the Why

In the present era, the presence of the internet has opened doors to a lot of possibilities. Talking about looking up video recipes to help you cook up a meal is the go-to option for the cooking amateurs, although those are premade videos that cannot assess the current state of your food and give you proper feedback. That's where our proposed product comes into the picture.

One of the primary focuses of the field of AI has been to make machines imitate human behaviour. Kochen Helfer is an approach to achieve this by using image processing and recommendation systems to equip an amateur with a personal cooking assistant.

2 Associated Publications

2.1 AI in Real-Life World

Artificial ıntelligence plays a significant role in our daily life. It is difficult to do anything which does not involve AI. With the advancements in technology, there has been a substantial impact on the businesses carried out. Strategies needed for any business to run smoothly are critical. This paper by Ghimire et al. [1], Prakash et al. [2] and Saxena et al. [3] has elaborately explained the growth of business and how machine learning and deep learning subsets of AI are used for tackling problems in business related to marketing, product recommendation, fraud detection and many more. The finance sector has also been significantly affected by the growing AI technology. The authors Zhang and Kedmey [4], Kolanovic and Krishnamachari [5] give a detailed discussion on the trends and attempts taken for processing business needs. The paper provides a model that can be used for the strategic modelling of an organisation.

Considerable use of AI in education has been witnessed in the last couple of years. Several tools have been created to make learning and teaching easy and interactive. Bhattacharya and Nakhare [6], Aldahdooh and Naser [7], Roll and Wylie [8] provide a detailed analysis of the use of AI-based tools in the vocational study sector and impacts it has on the students.

This pandemic situation made us realise that AI has played a significant role in dealing with problems, helping to build proper strategies for controlling the situation, suggesting appropriate plans, understanding the nature of the virus and generating

necessary medicines and vaccines. The paper by Nirmala and More [9] provides an insight into the importance of AI in fighting COVID-19.

2.2 AI in Smart Kitchen Appliance

In recent times, the rise in artificial ıntelligence has revolutionised our ways and standards of living. As the field expands, we see numerous applications of it in our daily life. One such application is in smart kitchen appliances. Be it a refrigerator, stove, storage system, AI finds its usefulness everywhere.

Some of these scenarios are discussed by Mallikarjun et al. [10], Floarea and Sgârciu [11], who have addressed the use of AI and IoT in an intelligent refrigerator. The paper proposed a method by which AI can help determine the quality and quantity of food present in the refrigerator, which can be notified to the user with the help of an android application. The paper also proposed a machine learning algorithm that would recommend recipes based on the fruits and vegetables present in the refrigerator.

Another such case is mentioned by Afroz et al. [12], who have proposed an AI-enabled gas stove with two-step safety and age verification. Basically, using an ML algorithm for age detection, it detects if a child is trying to turn on the stove and prevents them from doing so. The authors have implemented a machine learning object detection algorithm and a deep learning architecture using CNN, for system execution.

Keeping track of your dietary consumption is a crucial aspect of a person's life. We can do that manually by diet journaling, but for the elderly, it becomes a little challenging to keep track of their dietary habits. To tackle that, Gerina et al. [13] and Achananuparp et al. [14] proposed a way to track a person's diet by measuring some air quality patterns in the kitchen using a deep neural network. This shows how far AI has reached and how greatly it can impact our lifestyle.

2.3 AI in Everyday Cooking

The role of artificial ıntelligence is primarily to reduce manual effort. And that is where, in the modern world, technologies like the Internet Of Things, machine learning and deep learning are being applied in varied domains like business, finance, sports and so on. Similarly, the same is being used in the cooking industry. Here we will specifically focus on the application of artificial ıntelligence in everyday cooking and the cooking industry.

There have been several improvements and developments in this domain, and there are multiple examples that have been discussed here. As discussed by Bień et al. [15], Papineni et al. [16] there have been several issues with generating text, but with the help of text-developing neural networks, the process has become more

accessible. That is where Bień et al. [15] is mentioning how the recipe text dataset could help in this. In this research work, the author describes how the recipe text data and semi-structured datasets are related to each other by using a RecipeNLG dataset. The final goal of the author is to generate new recipes based on the recipe dataset provided to the model to learn. The author mentions that a cooking recipe consists of ingredients, quantity, and their unit of measures that could be used to train the model to extract food ingredients from a set of tokens. Now, once the model can identify food ingredients from a group of tokens, the same information is used to suggest new recipes that could be cooked.

Similarly, another such use case by Wang et al. [17] known as Market2Dish is described in their research work. It is a health-aware food recommendation scheme that maps ingredients sold in the nearby market to interesting, healthy dishes that could be cooked at home. This product primarily has three components, i.e. recipe retrieval, user-health profiling, and health-aware food recommendation. Each one of them has its work. Starting with recipe retrieval, it scans the ingredients available with the food consumer and based on that; it searches a large-scale dataset of recipes. Further, user-health profiling is done by capturing some parametric health numbers entered by the user and consumer itself. And finally, based on the health-aware food recommendation component, it provides the best healthy foods based on the input from the user.

3 Value Perspective

3.1 Need of the Product—A Helping Hand

A prevalent instance in a country like India is that after completing his schooling a student goes to some other city for his graduation where he lives all alone in a rented flat. He cannot have a maid at home as he gets lesser pocket money to run his days every month, and because of that, he needs to cook by himself.

The above narration is a pervasive story across thousands of students in India. It's easy for a student staying at home because he has his parents to take care of him. Here, the situation calls for a solution in the form of an assistant that could help the student in need. That is why we propose an AI-based cooking assistant that could help the student in the hour of help needed.

Kochen Helfer is a cooking assistant that works on the concepts of deep learning and natural language processing. there are two pieces to this entire product. The first piece is the generic cooking assistant that helps with the recipe. This assistant allows freshers in the domain of cooking by being constant support throughout the process. Text to speech conversion is needed here for the assistant to work. For example, let's say that someone wants to cook white sauce pasta. So, the person can go to the app and select the recipe he likes to cook. He can simply activate the assistant and ask to guide him with the recipe. Now, for cooking a white sauce pasta, one needs

to aggregate all the ingredients first. Pasta, along with all the spices, is collected, and then the process of cooking starts. Now, throughout the process of cooking, the assistant can give instructions for cooking the same. It can tell when to boil the pasta and prepare the sauce and finally mix the same to get the final output, i.e. the white sauce pasta (Figs. 1 and 2).

Now, the second part of the process or the app is the cooking correctness checker. This piece of the app works on the concepts of deep learning. To be specific, convolutional neural network can be used here. The specialty of this module is that it tells if your food is correctly cooked or not. The classification is done on the basis of stating the foods as undercooked, perfectly cooked and overcooked. This could be a

Fig. 1 White sauce pasta

Fig. 2 Ingredients to cook white sauce pasta

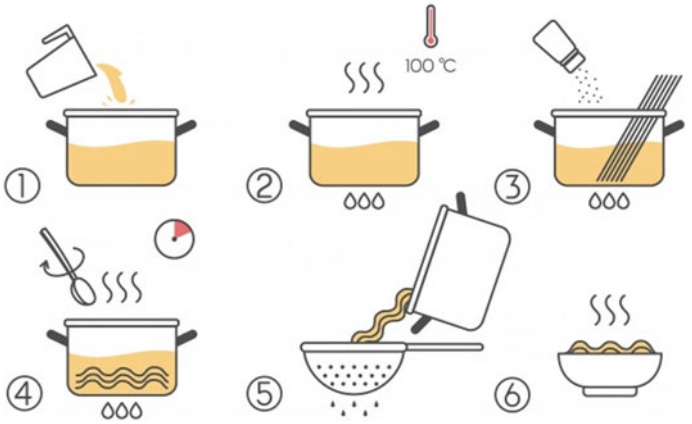


Fig. 3 Process of cooking white sauce pasta

perfect app to know if your food is cooked well. This works on a CNN model that could be trained on thousands of images taken while various recipes were executed (Fig. 3).

3.2 User Persona

The proposed product, Kochen Helfer, is designed to assist amateurs, who are new to the cooking world, and those who have experience in cooking. For amateurs, it can help them by giving them constant feedback about the state of the food, whereas, for experienced cooks, it can provide them with new recipes from the recipe book. From college students living in the hostel to the little kids aspiring to become a Masterchef or to the adults trying to pick up a hobby, anyone who wants to step into the cooking world or wants to learn something new will find this product helpful. The model building process follows the preprocessing technique.

4 Proposed Architecture

This section explains the visualisations available for viewing categorical and numerical data. There are separate visualisations for each of these broad categories. The different visualisations are shown below.

Fig. 4 High-level architecture

4.1 High-Level Architecture

The initial and most crucial part of the Kochen Helfer application is data requirement and gathering. We must have the correct data that is required. Once we have gathered the data, it becomes necessary to understand that the data is trustworthy. Preprocessing is the next step that comes into the picture after the Data Sanity check; here various processing techniques are applied to the data. Prepossessing is followed by the model building process. Testing of the application is done once the model is built. Deployment is the final process of creating the application (Fig. 4).

4.2 Data Requirement and Gathering

The requirements of the data for this product are:

1. The dataset should contain food images.
2. The images should be of the different stages of a recipe.
3. For each stage of a recipe, the dataset should contain images of uncooked, overcooked and perfectly cooked food.

For example, if we are cooking a curry, then the dataset should contain images of all the cooking stages, be it sauteing the vegetables or adding the spices. Moreover, for each stage, let's say "sauteing the vegetables", there should be enough images of un-sauteed, burnt and perfectly sauteed vegetables.

One such dataset resource can be the Recipe1M+ dataset which is the most extensive publicly available food dataset.

4.3 Data Sanity and Preprocessing

Once we have collected the image data of various dishes and their cooking stages, it is crucial for us to check if the corresponding data is correct. Some cooking experts can testify and attest to a few samples for their genuineness and correctness in this step. Such activity can help us conclude that the data collected for the model building process is correct.

For example, if we are trying to cook some rice-oriented dish, the domain expert can check and tell based on the images if the rice is overcooked, undercooked or

perfectly cooked. If the rice is overcooked, it breaks very easily and creates a slimy appearance. While if the rice is undercooked, then it does not shine white, and finally if the rice is cooked perfectly, then it shines white and grows big from its original size. Hence, such distinguishing factors can help deep learning models classify if the food is cooked correctly or not.

Going further, the preprocessing techniques that can be applied to the images for structuring the same are as follows:

1. ***Transformations*—**Such techniques are used to correct distortions or perspective issues in images. There are two types of transformations, affine and non-affine. Affine transformations include scaling, rotation and translation, while non-affine transformations tend to maintain collinearity and ıncidence. Non-affine transformations does not preserve parallelism (Fig. 5).

2. ***Rotations*—**This action rotates an image and places it in the correct order. A rotation matrix is used for the same. The angle of rotation is the angle by which the rotation has to happen (Fig. 6).

3. ***Scaling and Resizing*—**Interpolation is used to construct new data points within the range of a discrete set of known data points. Such techniques help to resize images, i.e. zoom in or zoom out as per requirements. Similarly, the Image Pyramid can be used to upscale or downscale images (Fig. 7).

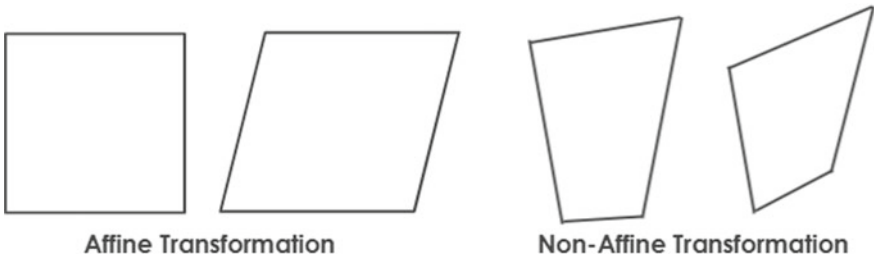


Fig. 5 Affine and non-affine transformations

Fig. 6 Image rotation

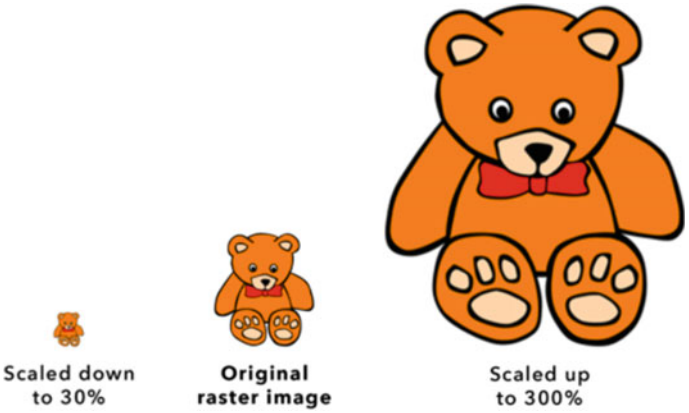


Fig. 7 Image scaling

Fig. 8 Image cropping

4. ***Cropping Images—***Several times, the entire image is not needed, and only a specific portion is required. In such cases, the images can be cropped, excluding the non-required area of the images (Fig. 8).

5. ***Convolutions and Blurring—***Convolution is a mathematical operation performed on two functions producing a third function, typically a modified version of one of the original functions. Usually, a Kernel is defined, which is an n × n matrix that can be run on our image set. Image blurring is an application of convolution where the pixels are averaged within a region (Fig. 9).

6. ***Sharpening—***A similar application of convolution is sharpening. Sharpening strengthens and emphasises the edges of an image. As this involves convolution, a kernel is again used. The sum of the kernel should be 1. If the kernel does not sum up to 1, then the image's brightness and contrast may change (Fig. 10).

7. ***Thresholding and Binarization—***Thresholding is an activity where an image is binarised. But before an image is binarised, the image has to be converted

Fig. 9 Image blurring

Fig. 10 Image sharpening

to grayscale. Usually, simple thresholding requires us to provide the threshold value, but Adaptive Thresholding auto manages the same and does not ask us to define it (Fig. 11).

8. ***Dilation and Erosion*—**Dilation is the process of adding pixels to the boundaries of objects in an image, while Erosion is the process of removal of pixels from the edges of the objects in an image (Fig. 12).

9. ***Edge Detection and Image Gradients*—**Edge detection is the process of capturing sudden changes in an image. This is a significant step as it helps in feature engineering for deep learning models (Fig. 13).

Fig. 11 Image binarization

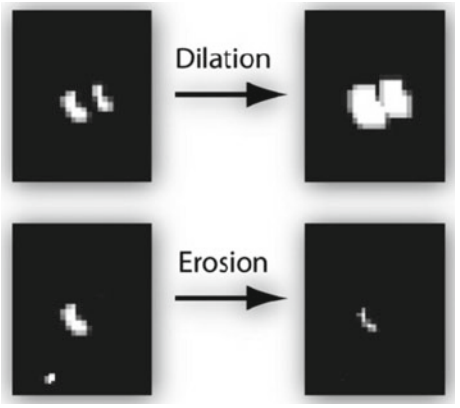


Fig. 12 Dilation and Erosion

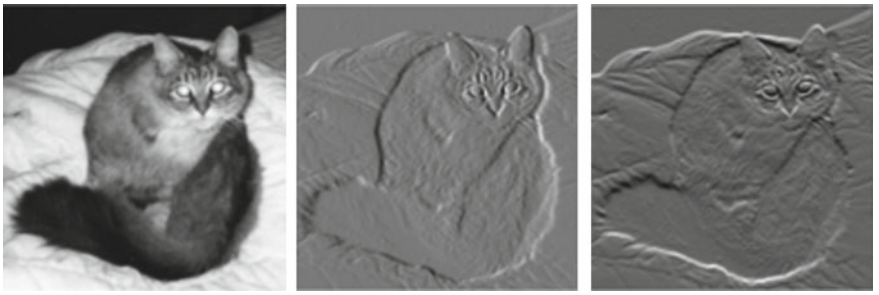


Fig. 13 Image gradient

4.4 Model Building

One of the most important steps in building an AI-enabled cooking assistant is food ımage processing, which involves showing many food images to a mathematical model that can then classify food images into various subclasses. The proposed product, Kochen Helfer, will categorise any phase of cooking into three categories, namely, "perfectly cooked", "undercooked" and "overcooked" (Fig. 14).

In recent times, convolution neural networks (CNNs) have given us a massive breakthrough in the field of Image analysis. CNN is a class of deep learning neural networks and is most commonly used to analyse visual images and classify them into different categories. A neural network is a system of hardware and/or software that is designed to operate in the way a neuron does in the brain. In CNN, the neurons are arranged more like those in the frontal lobe of the brain, which is an area that deals with the visual stimuli. A CNN is a multilayer system designed in a way to reduce processing. The different layers of a CNN are as follows:

1. Convolution Layer
2. Pooling Layer
3. Fully Connected Layer
4. Normalisation Layer

Unlike any ML algorithm, where we need to show features to the ML algorithm manually, CNN extracts features from the pixels of the image and convolves them into a much simpler set of machine-understandable features. These features can then be used to identify and classify any image that is passed through the network. Training a CNN needs a lot of pictures and thus takes a lot of processing time. Therefore there are many CNN architectures available (e.g. AlexNet, GoogleNet) with pre-trained weights, which can then be fine-tuned and used for any problem statement according to need. This gives better accuracy since the architecture is pre-trained on many images to understand the features of an image and distinctly classify them.

Some examples are Yadav et al. [18], Zhou et al. [19], Simard et al. [20] where SqueezeNet and VGG-16 CNNs are used to classify food images automatically. A similar kind of architecture with some fine-tuning to classify images into "perfectly cooked, undercooked and overcooked" can be used for the proposed product, Kochen Helfer.

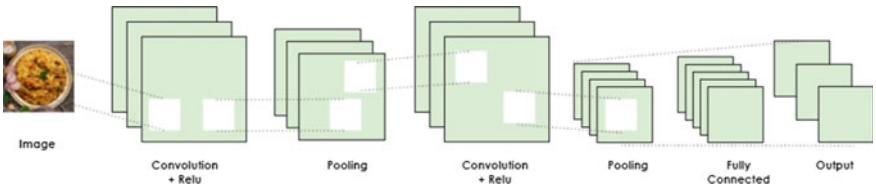


Fig. 14 Process of convolutional neural network

4.5 Model Testing

Testing plays a very vital role in any development process. It is the crucial step that decides whether the model is ready to use or requires further development. Hence it is essential to follow the right path for choosing the best tool for testing and creating the test cases.

Testing needs to be done end to end. So it is essential to cover all the scenarios.

These typically include functional testing, usability testing, performance testing, fit and finish testing, regression testing, device specific testing and user acceptance testing.

Below are some points that need to be kept in mind while selecting the test cases:

- Multiple scripting languages should be supported
- The application is going to be deployed in multiple mobile applications, so the creation of test scripts in various languages should be possible.
- Integration of the testing tool with the CI/CD pipeline.

In recent times an automated mobile application testing tool named Appium is a popular tool that can be used to test hybrid or native iOs and Android applications. Appium uses the WebDeriver interface to run the test cases. The reusability of code features for iOS makes the tool robust. Integration with the CI/CD tool is simple.

4.6 Model Deployment

Software deployment can be considered as a combination of processes that makes a software system available to the user experience. Deployment of the webserver into a scalable production environment and deployment to the play store or apple store are the two main components of a deployment process. The webservers help to transfer data to, and fro from the app; hence proper configurations are essential. Similarly, for the app store, we will be required to provide screenshots, marketing material and fill out forms of various stores mobile enterprise application.

Mobile Enterprise Application Platform (MEAP) as middleware platforms can be used to achieve a high level of flexibility. They provide quick deployment procedures with the provision of high-level languages. Most importantly, they help deploy the application once and can be deployed in various mobile device types.

5 User Interaction Interface

5.1 Log In View

The Log In View is the application's landing page. The main purpose of this view is to get the user registered into the application if it's a new user or gets the user logged into the application for an already registered user. The view contains a text field for the username and another for the password, along with the login button. Apart from that, there is a register button to get the new users signed up. Pressing the register button opens up a registration form where the new users can enter their name, age, email ıd, phone number and set up a password. After getting registered, the user can log in using their credentials, and on pressing the login button, they are redirected to the homepage view (Fig. 15).

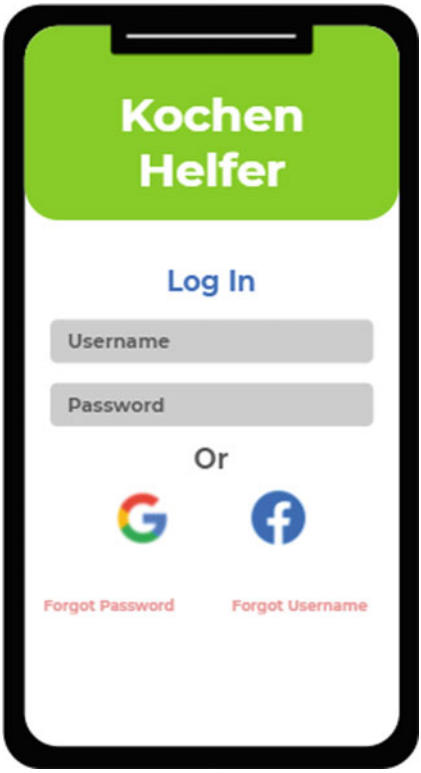


Fig. 15 Login view

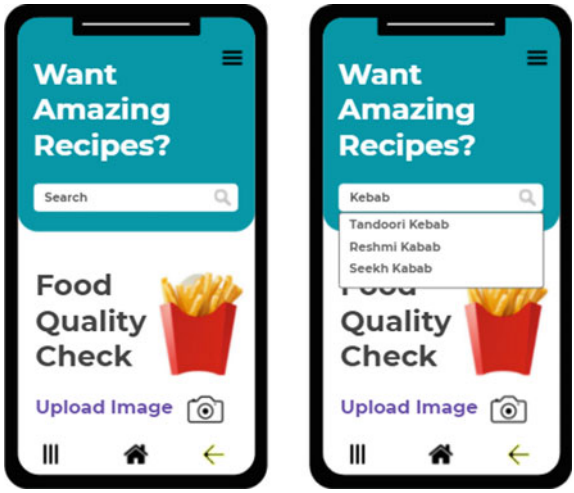


Fig. 16 Home page and recipe selection view

5.2 Homepage View

The homepage view is the view that you land on after logging into the application. The primary purpose of this view is to encapsulate all the different features present in Kochen Helfer and to provide a user-friendly ınterface between the user and the functionalities.

This view contains a side panel that contains a menu for all the different functionalities. Next, we have a button to take us to the recipe selection and uploads view. The view also contains a recipe book using which the user can go through any online recipe (Fig. 16).

5.3 Recipe Selection and Uploads View

The recipe selection page provides you with the different cuisines that can be made with the help of the application. It contains recipes of food from various regions segregated categorically. Once we click on a particular recipe, we are then directed to the instructions that need to be followed for the cuisine. Each step is explained in a very detailed manner.

The upload view feature of the application provides the user to upload the current status of the food being prepared. The primary purpose of this page is to make the experience really interactive and provide a comfortable cooking experience.

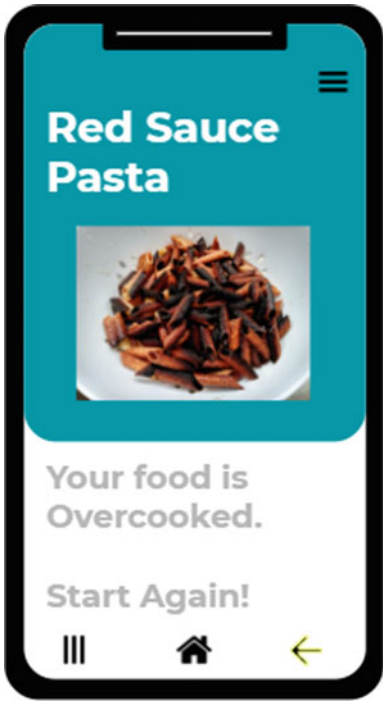


Fig. 17 Results view

5.4 Results View

In the view result section, the evaluation results of the uploaded image are displayed. The interface helps to understand if the particular stage of the food is cooked or if it needs time. Accordingly, it displays "undercooked", "cooked" and "overcooked". If it is cooked or overcooked, the next step in the process is displayed on the screen. This helps the user to track their progress in every step (Fig. 17).

6 Conclusion

As we can see, the proposed approach can help amateurs in cooking while being away from home. Such a product could develop independent individuals who can take up their responsibility. With time, such applications could help amateur individuals with very minimal knowledge of cooking to have expertise in the same.

With the right model in the right place, it could provide quality insights into whether the food being cooked is correctly cooked or not. It could save millions of people from wasting food, given that they have a guide to help them out.

7 Future Work

This project is currently a theoretical concept. The real implementation of this project is yet to take place. In the near future we will be working on the application of the entire Kochen Helfer for machine learning to mobile application building.

References

1. Ghimire, A., Thapa, S., Jha, A.K., Adhikari, S., Kumar, A.: Accelerating business growth with big data and artificial ıntelligence. In: 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 441–448 (2020). https://doi.org/10.1109/I-SMAC49090.2020.9243318
2. Prakash, S., et al.: Characteristic of enterprise collaboration system and its implementation issues in business management. Int. J. Bus. Intell. Data Min. **16**(1), 49–65 (2020)
3. Saxena, A., et al.: A review of clustering techniques and developments. Neurocomputing **267**, 664–681 (2017)
4. Zhang, X.P.S., Kedmey, D.: A budding romance: finance and AI. IEEE MultiMed. **25**(4), 79–83 (2018). https://doi.org/10.1109/MMUL.2018.2875858
5. Kolanovic, M., Krishnamachari, R.T.: Big data and AI strategies: machine learning and alternative data approach to investing (May 2017)
6. Bhattacharya, P., Nakhare, S.: Exploring AI-enabled ıntelligent tutoring system in the vocational studies sector in UAE. In: 2019 Sixth HCT Information Technology Trends (ITT), pp. 230–233 (2019). https://doi.org/10.1109/ITT48889.2019.9075093
7. Aldahdooh, R., Naser, S.S.A.: Development and evaluation of the oracle ıntelligent tutoring system (2017)
8. Roll, I., Wylie, R.: Evolution and revolution in artificial intelligence in education. Int. J. Artif. Intell. Educ. **26**(2), 582–599 (2016)
9. Nirmala, A.P., More, S.: Role of artificial ıntelligence in fighting against COVID-19. In: 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE), pp. 1–5 (2020). https://doi.org/10.1109/ICADEE51157.2020.9368956
10. Mallikarjun, B.C., Harshitha, S., Harshita, B.K., Bhavani, S., Tarwey, S.: Smart refrigerator: an IOT and machine learning based approach. In: 2020 International Conference for Emerging Technology (INCET), pp. 1–4 (2020). https://doi.org/10.1109/INCET49848.2020.9154025
11. Floarea, A.-D., Sgârciu, V.: Smart refrigerator: a next generation refrigerator connected to the IoT. In: International Conference—8th Edition Electronics Computers and Artificial Intelligence (2016)
12. Afroz, M., Hasan, N., Hossain, M.I.: IoT based two way safety enabled ıntelligent stove with age verification using machine learning (2021). https://doi.org/10.1109/ICCCI50826.2021.9402701
13. Gerina, F., Massa, S.M., Moi, F., et al.: Recognition of cooking activities through air quality sensor data for supporting food journaling. Hum. Cent. Comput. Inf. Sci. **10**, 27 (2020). https://doi.org/10.1186/s13673-020-00235-9
14. Achananuparp, P., Lim, E., Abhishek, V.: Does journaling encourage healthier choices? Analyzing healthy eating behaviors of food journalers. In: Kostkova, P., Grasso, F., Castillo, C., Mejova, Y., Bosman, A., Edelstein, M. (eds.) Proceedings of the 2018 İnternational Conference on Digital Health, ACM, pp. 35–44 (2018)
15. Bień, M., Gilski, M., Maciejewska, M., Taisner, W., Wiśniewski, D., Ławrynowicz, A.: RecipeNLG: a cooking recipes dataset for semi-structured text generation (2020)

16. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318 (2002)
17. Wang, W., Duan, L.-Y., Jiang, H., Jing, P., Song, X., Nie, L.: Market2Dish: health-aware food recommendation. ACM Trans. Multimed. Comput. Commun. Appl. **17**, 1–19 (2021). https://doi.org/10.1145/3418211
18. Yadav, S., Alpana, Chand, S.: Automated food image classification using deep learning approach. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 542–545 (2021). https://doi.org/10.1109/ICACCS51430.2021.9441889
19. Zhou, L., Zhang, C., Liu, F., Qiu, Z., He, Y.: Application of deep learning in food: a review. Compr. Rev. Food Sci. Food Saf. **18**, 1793–1811 (2019)
20. Simard, P.Y., Steinkraus, D., Platt, J.C.: Best practices for convolutional neural networks. In: 12th International Conference on Document Analysis and Recognition, vol. 2 (2003)

Contemporary Digital Consumers: "Real" Versus Virtual Value Co-creation

Milena Ratajczak-Mrozek and Marcin Wieczerzycki

Abstract Digital technologies are used extensively by both producers and consumers. While this makes it easier for both types of actors to interact and co-create value, increasing the prevalence of this process, the very fact of mediating an interaction through a particular type of technology (digital media in this instance) influences its character as well. Therefore, the purpose of the chapter is to conceptualise the characteristics of digital co-creation within the theoretical framework of Service-Dominant Logic (SDL). We do this by contrasting digital co-creation with co-creation occurring in the context of the "real", unmediated world.

We contribute to SDL literature by suggesting five characteristics of digital value co-creation, distinguishing it from co-creation occurring in the unmediated context: co-presence, automatization, simulation, antagonism and playfulness. We illustrate these traits with real-life examples. Additionally, we conceptualise how these characteristics are linked with the most important traits of modern digital consumers.

1 Introduction

Digitalization is a phenomenon that to some extent affects every person. We all are influenced directly or indirectly by "the use of digital technologies" ([25], p. 79) which is characteristic of digitalization. These technologies include social media, big data analytics, and cloud and mobile technology, which are globally used by consumers and/or companies alike. Therefore, most consumers' consumption of products and services is no longer limited to the "real", unmediated world, but also occurs in a virtual context, where all the actions and interactions are mediated by digital technologies. Thus, they become digital consumers who purchase products online or take advantage of digital content or search for information online [26]. Such modern digital consumers have access to many media channels, communicate

M. Ratajczak-Mrozek (✉) · M. Wieczerzycki
Department of International Marketing, Poznań University of Economics and Business, Poznań, Poland
e-mail: milena.ratajczak-mrozek@ue.poznan.pl

S. S. Rautaray et al. (eds.), *Data Science in Societal Applications*, Studies in Big Data 114, https://doi.org/10.1007/978-981-19-5154-1_10

extensively with each other as well as with companies, and are interconnected. These traits, combined with a wider choice of products and services and higher expectations, as well as reciprocity and interconnectedness between digital and the unmediated context of consumers' lives, result in the empowerment of modern digital consumers. In turn, the digitalization and empowerment of digital consumers influences how the value of companies is co-created.

Value, understood as a trade-off between the benefits and sacrifices that occur during the process of interactions [11, 33], is the main aim of companies' activities [6]. The concept of value co-creation is mostly linked with Service-Dominant Logic [27, 29], seeing every economic activity as a service-for-service exchange [29]. In SDL "value is co-created by many actors, including consumers" ([29], p. 47). The process of co-creation should end in positive outcomes [27], although one needs to be aware that destruction or diminishment of value is also possible [20].

Taking into consideration the unavoidability of digitalization and the fact that the boundaries between real (unmediated) and virtual (digital) contexts tend to penetrate each other and overlap [9], important questions arise: how is value co-created in the digital context, and what is the difference between unmediated and digital value co-creation? To the best of our knowledge, these questions have not been raised in the current literature. While the importance of digital technology is often noted, the research itself is mostly focused on the problem of the empowerment of modern digital consumers [9, 21]. Taking into consideration the indicated research gap, the aim of the chapter is to conceptualise the characteristics of digital co-creation. We do it by comparing digital co-creation to the co-creation in the "real", that is unmediated, context. We understand this "real" (unmediated) co-creation as all the interactions that occur directly and without the help of digital technologies, in contrast to virtual (digital) co-creation, where these technologies serve as an intermediary between interacting actors.

Although we focus on positive outcomes and value co-creation, we also discuss possible negative outcomes of consumers' involvement in value creation processes, that is value co-destruction.

In the chapter, we apply the conceptual review method [15] to propose a conceptual framework of characteristics of digital value co-creation. The result of applying a conceptual review is "a theoretical contribution that refines, reconceptualizes, or even replaces existing ways of viewing a phenomenon" ([15], p. 27). We analysed the latest literature in the area of digital consumers and digitalization. Based on our own conceptual work, we suggest five characteristics of digital value co-creation and co- destruction resulting from the empowerment of digital consumers.

We contribute to the up-to-date literature by suggesting five characteristics of digital value co-creation, distinguishing it from co-creation occurring in the unmediated context. These characteristics are: co-presence, automatization, simulation, antagonism and playfulness. We provide real-life examples of how these traits influence value co-creation. Additionally, we conceptualise how these characteristics are linked with the most important traits of modern digital consumers. The links between the individual characteristics of digital value co-creation and possible directions of further research are also discussed.

The remaining part of the chapter is structured as follows. First, we define modern digital consumers and identify the six most important traits characterising these consumers. Second, we discuss Service-Dominant Logic and the resulting concept and specifics of value co-creation. Based on these analyses, in next section of the paper, we conceptualise the five most important characteristics specifying digital value co-creation and propose the conceptual framework of the characteristics of digital value co-creation resulting from the empowerment of digital consumers. Our chapter finishes with managerial implications and indications of directions for further research.

2 Digitalization and Digital Consumers

Digitalization means "the use of digital technologies" ([25], p. 79), which includes amongst others: cloud and mobile technology, social media, big data analytics, the Internet of Things, virtual (augmented) reality, cognitive technology and security [25]. This global phenomenon has a tremendous impact on people and consumers all over the world. Digital technologies create new behaviours and communication patterns between consumers and companies. Therefore, the debates about e-consumers and digital consumers are very timely and take on importance. Some authors, like [9, 33], even discuss the idea of digital consumer culture which is "shared sets of consumption behaviour that directly or indirectly emanate from people's interactions with digital technologies, such as the Internet, social media, mobile devices and applications" ([9], p. 2). An e-consumer is a person that buys products online. A digital consumer in turn may be e-consumer purchasing products online, but also a person that takes advantage of digital content published online (e.g. watching YouTube) or searches for information online. Therefore, the term digital consumer is broader than just e-consumer [26].

There are several traits that characterise modern digital consumers. These include

1. Internet usage and having access to numerous media channels,
2. information-empowerment,
3. accelerated communication (between consumers and companies),
4. interconnectedness,
5. reciprocity between digital and unmediated contexts of consumers' own lives,
6. wider choice of products and services linked with higher expectations and requirements.

First of all, digital consumers are much more engaged with digital technologies, which especially include the Internet, social media, and "on-demand" media [9, 26]. This is also related to the fact that they have continuous access to numerous media channels, which also include different so-called "digital engagement platforms", such as social media sites and apps [19].

Continuous access to different digital media channels is directly linked to the second trait of digital consumers, namely information empowerment [1, 9, 21].

Modern digital consumers have access to many sources of information. At the same time, in the digital context consumers are not only recipients of information but also active creators of information resources [21]. This access to information brings not only benefits such as the capability of making more conscious decisions, but also some flaws. These include information overload, confusion, and difficulty to distinguish between true and false news. The challenge is to discern useful information with minimal time, effort and energy [9].

Access to numerous digital media channels, especially the Internet and social media, facilitates communication. Digital technologies have changed the way consumers communicate with each other and make decisions [21]. Thanks to digitalization, consumers not only communicate with each other, but now it is easier than ever before to communicate directly with companies and therefore influence their activities [17]. The accelerated communication allows increased clustering and the building of consumer communities around brands which foster their engagement [18].

All the traits mentioned above result in the much higher interconnectedness of modern digital consumers compared with consumers operating only in the unmediated context [1, 9]. As ([9], p. 1) underline: "The convenience and connectedness provided by social media, mobile technology and other forms of digital technologies and applications promote assimilation, integration or acculturation beyond the users' 'own community'". This interconnectedness and possible interactions between consumers reach far beyond geographical, national or ethnic borders. Digital media and technologies allow consumers to overcome the limitations of time and space, and therefore they can interact in both synchronous and asynchronous ways. The interconnectedness between digital consumers has far reaching consequences. On the one hand, digital technologies foster social interaction [18], which results in the empowerment of consumers [9] who can connect around a similar cause. It is easier than ever before to find similarly minded people. On the other hand, social media segregates and divides people. For instance, content on social media often happens not to be neutral or politically correct [9].

Although we focus on the most important traits of digital consumers, we have to underline the reciprocity between the digital and unmediated contexts of the lives and existence of these consumers [9]. Digital consumers' activities are not limited only to the digital environment, and the two (that is digital and unmediated) contexts tend to intermingle, "as often the online and offline boundaries are blurred and overlapped" ([9], p. 2). For example, a consumer can buy a product online but consume it offline, or they use a traditional unmediated service but rate it later online as part of post-purchase evaluation. Moreover, because of social media and the constant sharing of our lives online (including check-ins, different statuses, virtual/augmented reality), these two contexts tend to overlap even more [9].

Finally, information empowerment, communication and the interconnectedness of modern digital consumers allow them to have a wider choice of products, but also higher requirements and expectations from the companies [32]. At the same time, all the above-mentioned six traits of modern digital consumers result in their empowerment [9, 21]. Empowerment, meaning "the act of giving somebody more

control over their own life or the situation they are in" ([19), is also used in the context of "empower[ing] individuals to take an active role in the design of new products and processes" ([21], p. 222), which is clearly linked with the creation and co-creation of value.

3 Service-Dominant Logic and Value Co-creation

The empowerment of consumers in both unmediated (real) and digital (virtual) contexts translates into their inclusion in the process of companies' value creation, that is value co-creation. In the current scientific literature, there is no single understanding and definition of value. This is linked to the fact that value is subjectively experienced and perceived [28]. In its broadest sense, it is understood as a trade-off between the benefits and sacrifices that occur during the process of interactions between different actors (entities) [11, 34].

The concept of value co-creation, although discussed extensively within different scientific streams and marketing schools (e.g. Industrial Marketing and Purchasing Group [6]), is mostly linked with Service-Dominant Logic (SDL) [27, 29]. SDL is a theoretical framework that posits that all exchange is based on services. Services are "re-conceptualised by abandoning the intangible-unit-of-output meaning they had acquired through the industrial-, production- and goods-dominant orientation" ([29], p. 47). Every economic activity is a service-for-service exchange [29]. It takes place through the integration of resources which leads to the co-creation of value [27]. The sources of competitive advantage are the so-called "operant resources", that is knowledge and competencies, and their implication [30]. The focus on operant resources in SDL changed the interpretation of the primary unit of exchange in economic activities (Brodie et al. 2019, p. 5), whereas in SDL goods are seen only as distribution or transmission mechanisms for service provision [27].

The most important assumptions of SDL concern value and value co-creation. SDL is opposed to the traditional view of goods-dominant logic (GDL). GDL assumes value-in-exchange, where value is created (produced) by companies and delivered to customers in exchange for money. Therefore, GDL sees producers and consumers as fundamentally different and carrying out vastly different roles. SDL takes the opposing view, seeing both producers and consumers as actors capable of creating value. In SDL, "value is co-created, rather than created by one actor and subsequently delivered" ([29], p. 47). Consumers are no longer seen as passive, and are actively involved in value co-creation. As ([30], p. 146) underline: "The roles of producers and consumers are not distinct, meaning that value is always co-created, jointly and reciprocally, in interactions among providers and beneficiaries through the integration of resources and application of competences". These consumers' and other actors' involvement in economic activities and business processes in the form of value co-creation is the source of companies' advantage [21].

The fundamental premises (FP, or later called axioms) of SDL have evolved over time. Nevertheless, currently, the most relevant axioms concerning value include [28,

29] : "Value is co-created by multiple actors, always including the beneficiary" (FP6), "Actors cannot deliver value but can participate in the creation and offering of value propositions" (FP7), "Value is always uniquely and phenomenologically determined by the beneficiary" (FP10), and "Value co-creation is coordinated through actor-generated institutions and institutional arrangements" (FP11).

Value co-creation focuses mostly on positive outcomes [22, 27]. However, it needs to be underlined that mere interaction with consumers does not guarantee the creation of value [3], and the whole process of interaction between actors can result in value co-destruction [10, 20, 21]. Value co-destruction is defined as "destruction or diminishment of value for one or more actors" [20], p. 2), or as a situation "when the interaction and its outcome are perceived as unfair or unsatisfactory and lead to dysfunctional attitudes in the actors who participate in the process" [21], p. 222). It can be summarised as win-lose or lose-lose outcomes of interaction [20]. It results, amongst other things, from the negative experiences and perceptions of actors [3], misalignment of the actors' practices, including consumer misbehaviour [10], misuse of resources [21], opportunistic behaviour, dishonouring contractual promises or other conflictual interactions [20].

It is also important to state that value co-creation and co-destruction can coexist when one actor notes positive outcomes of value co-creation and the other one mostly notes negative ones (that is value co-destruction) [10, 20].

4 Specificity of Virtual Value Co-creation

4.1 Co-presence

The first characteristic of digital co-creation is the so-called consumer co-presence. This term denotes a situation where the exchange between the seller and the consumer is influenced by interactions with other consumers, or an otherwise uninvolved audience (Colm et al. 2017). While this phenomenon can, and indeed does in some cases, occur during unmediated value co-creation (for instance when patients chat with each other in a clinic's waiting room), it is much more prevalent in the digital environment, thus becoming a distinct characteristic of the digital value co-creation process.

Due to the particular ability of digital media to overcome the limitations of time and space, leading to their compression, countless consumers can occupy the same virtual places, interacting in both synchronous and asynchronous ways, and thus create a state of permanent consumer co-presence, increasing the number of potential co-creational processes occurring at any given time. The most obvious example involves internet reviews. While gathering information on the best restaurant in the neighbourhood would normally entail asking around and require some time, digital consumers can access the opinions and experiences of others at any moment. In many cases relying on co-presence is not even a conscious decision. For instance, products in an online shop might be sorted according to their review scores, which means

that co-presence can impact the value co-creation process without the consumer's knowledge.

Moreover, due to the nature of digital media, a large part of interactions between the company and its consumers (like those taking place in comment sections on social media, discussion boards, etc.) is recorded and visible to all the other actors. Therefore, even when consumers do not decide to consciously share their experiences in the form of reviews, these experiences still can frame and influence the experiences of other consumers.

This has some consequences for the character of the value co-creation process. On the one hand, it becomes more history- and context-dependant. On the other hand, the context itself becomes significantly broader. This matches well with the more recent trends in SDL literature, where the old approach of focusing on a singular producer-consumer dyad gives way to the studies of entire constellations of such relationships [4], as well as institutions that emerge from them and regulate them [28]. However, this can sometimes leave the company with less time for in-depth individual interactions with the consumers, since at any given moment there is the potential for co- creational processes occurring, each involving more actors. In some cases, this creates the need for automatization, which will be discussed in the following section.

For the outcomes of the value co-creation process, the broader context can have both positive and negative impacts. On the one hand, due to the involvement of more actors, sometimes there is less strain on the company to contribute to the process, since some actions can be performed by other consumers. On the other hand, however, this comes with diminished control over the entire process. A good example of co-presence being both a blessing and a curse in the value co-creation process can be found during streaming events. Many streams rely on the interactions between the streamer and the chat, which consumers use to communicate (which can take the form of a simple conversation, banter or some sort of a game). Therefore, chat becomes an essential component of the consumer experience. However, since streamers only have limited control over what is being said or done in the chat, sometimes it can lead to problems – for instance when the use of, e.g. racial slurs in the chat negatively impacts the streamer's own image.

4.2 Automatisation

Another unique feature distinguishing digital co-creation from the traditional one is the extent to which the interactions between different actors participating in the process are automatized. In the case of unmediated co-creation, the process largely depends on individual interactions between actors. For instance, the individual consumer's experience with a hairdressing service is shaped during their interaction with a hairdresser, with both sides exchanging information in a way that allows adjustments of the service to the needs of the consumer, leading to the co-creation of in-use value. This interaction can be formalised to some degree (e.g. with the use of

forms). However, usually at least some amount of interaction between actors themselves is required. Digital co-creation, however, can in many cases completely omit this stage, relying instead on the mediation of technological interfaces, which allow consumers to customise their experience without the need to ever directly interact with other human beings [16]. Think of computer hardware stores, where consumers can use the website to configure their rig out of available components, which will then be built by the employees and delivered directly to the consumer. While this process still constitutes an act of value co-creation, and the consumer receives his own personalised experience, the whole endeavour plays out with no direct human interaction whatsoever.

Sometimes, this reliance on technical interfaces and automatization can be purposefully obscured in the process of value co-creation. In the above-mentioned example of configuring PC hardware, the consumer is aware that they are using a tool rather than communicating with a human being. However, this fact can also be hidden from the consumer—for instance during communication via email or a Facebook chat. In similar situations, the consumer can receive fully customised messages, constructed using a number of templates and filled with consumer data. While such an effort can be transparent to more technologically savvy consumers, others may believe that they are interacting with an actual human being. At the same time, should the consumer's inquiry not match any of the predetermined templates, communication with a company's employee may be initiated. This allows for more efficient value co-creation with the consumers while avoiding the pitfalls of typical mass communication, which makes it harder to generate meaningful experiences for consumers.

Sometimes, this automatized way of value co-creation is built into the product itself. For instance, music streaming services like Spotify track the songs and bands that consumers are listening to, which is then used to suggest new artists that are similar in style, and should therefore be compatible with consumers' tastes. Again, no direct human-to-human interaction is required, but this time even the consumer's input has been automatized and does not require them to consciously engage in communication with technical interfaces. This allows companies to optimise the value co-creation process even further.

4.3 Simulation

The important characteristic of co-creation taking place within an unmediated reality is its grounding in real interactions in a specific time and place between real people. This can lead to the development of a genuine relationship between the producer and the consumer. For instance, a hairdresser may remember the consumer's name and preferences and have some insight into aspects of their life, which can fuel meaningful interactions transcending the boundaries of a market exchange, and which contribute to the process of co-creating a consumer's experience. Such interactions

are particularly important for consumers in the context of loosening social ties and the growing problem of loneliness and feeling disconnected.

Companies interacting with consumers digitally face the challenge of replicating these interactions within the technologically mediated environment. However, despite many digital interfaces offering tools for personalised communication, due to its scale and simultaneousness, it essentially remains mass communication. Moreover, communication through digital media lacks some of the social cues that allow the contextualisation of consumer experience (the so-called social presence; [23] and make it truly unique and memorable, which hurts its co-creational potential. For instance, while an unmediated interaction is informed and contextualised by facial expressions, body language, tone of the voice, as well as physical surroundings, the interactions occurring in digital media oftentimes are limited to the exchange of text and image-based messages. This results in a lower overall feeling of physical and emotional proximity with the partner during the interaction [23].

To counteract this limitation of digital communication and foster meaningful co-creational experiences for the consumers during interactions, companies rely on the process of simulation to artificially produce social cues and context where there are none, creating what philosophy refers to as simulacra [2]. A simulacrum is an imitation, which in the process of signifying something else becomes its own thing and eventually cannot be distinguished from the original (becoming a part of hyper-reality). [2] described this process using the example of simulating an illness. If in the process of simulation, one manages to trigger the same symptoms that are caused by the illness, then the simulacrum of illness becomes a part of reality to the same extent as the illness itself, and the two become effectively indistinguishable. Thus, the understood process of simulation has always been a vital part of marketing endeavours, becoming a subject of theoretical elaboration by marketing scholars, particularly those from the Consumer Culture Theory (CCT) literature stream [12, 13, 31]. However, while simulation has always been utilised to project meanings onto products and services to turn them into symbols (e.g. Apple products becoming simulacra of creativity, artistic nature, etc.), in the digital environment this process occurs on a much larger scale and encompasses new areas.

Most importantly, simulation becomes an answer to the previously described problem of low social presence characteristic of digital media. Companies use it to try and infuse brand profiles on digital media with different personalities, thus humanising them, creating simulacra of real people. This contributes to the consumer's feeling of interacting with an actual human being, which diminishes the detrimental effects of the low social presence of digital media. For instance, Wendy's Twitter profile is notorious for its brash communication, oftentimes poking fun at both consumers and competitors. Conversely, Denny's Twitter profile builds its personality around sharing memes and partaking in *non-sequitur*-filled conversations with consumers. These clearly defined personalities provide consumers with co-creational experiences which would otherwise be difficult to evoke.

That being said, while the simulation allows companies to overcome the limitations of the digital environment, it is still incapable of providing a perfect substitute for a genuine unmediated interaction. While consumers may communicate with

brand social media profiles as if they were their friends or at least acquaintances, such interactions lack the depth and substance of real-world interactions. Consumers may develop some sort of emotional attachment towards the brand, but these feelings will remain unreciprocated, as the brand profile in question remains a carefully crafted fictional character. And even though it could be argued that the social media managers behind these profiles could develop relationships with these consumers, this remains highly unlikely due to the sheer number of these interactions (which is not limited by time and space, as is the case for face-to-face interactions).

Moreover, while these interactions with brand simulacra can provide consumers with unique experiences, thus leading to value co-creation, there are situations when the simulation fails, and the attempt itself comes to light, which can become a source of conflict leading to value co-destruction. For instance, when the Twitter profile of SunnyD tweeted "I can't do this anymore", and other brands, such as Little Debbie, entered the conversation as if it was a real person going through a crisis, offering tips and emotional support, many consumers saw it as brands simulating depression for marketing purposes and found it inconsiderate or downright offensive.

4.4 Playfulness

Another key characteristic of value co-creation occurring via digital media is its playful character. This is another direct consequence of digitalization and mediation itself, with the internet and social media influencing the way consumers and companies communicate and interact, thus impacting the overall experience. The common challenge faced by actors utilising said media is the overabundance of available information (commercial and otherwise), which leads to cognitive overload and makes attention a key resource, sought by companies and consumers alike—leading to the development of concepts such as attention economy [8, 14, 24]. This means that, on the one hand, companies need to compete with each other for consumers' attention, and, however that consumers themselves will try to capture the attention of the company and fellow consumers.

This changes how the communication is carried out. Since there is too much information for anyone to comprehend, the actors strive to distinguish themselves from all the others to capture as much attention as possible. Therefore, to-the-point and generic business communication quite often do not provide satisfactory results, and actors instead try to introduce some playful elements into the interaction, to make it more interesting and engaging, and thus capable of capturing more attention (and keep it for longer). This can include using less formal language, or incorporating jokes, wordplay or Internet memes into communication. Playfulness can also lead to more elaborate interactions involving different actors. For instance, when fans of the DotA 2 video game tried to persuade its producer, Valve, to organise the annual game event called Diretide, their efforts took a strange turn, and consumers ended up posting their demands on the fan page of another brand with a similar sounding name—Volvo. These messages were so numerous that Volvo decided to respond.

However, rather than simply moderating the communication and deleting all the comments that had nothing to do with their business operations, the brand decided to play along, and actually petitioned Valve to organise the event in question in a joking manner. This support for the consumers' cause from a company that seemingly had nothing to do with the entire situation and was included as a form of soft trolling created a unique experience for consumers, and met with their positive response.

However, this playfulness can sometimes become a hindrance to the process of value co-creation, or even lead to value co-destruction. When consumers act in a search for the attention of their fellow consumers, their goals do not necessarily align with those of the company. This is particularly evident in the many cases of failed attempts at crowdsourcing made by companies, where consumers make suggestions that are meant as jokes (sometimes highly inappropriate and offensive ones), rather than treating this interaction seriously. For instance, when BC Ferry Services asked consumers to name its new ferry via an open poll, the consumers proposed a myriad of suggestions that were clearly poking fun at the company, such as "S.S. ShouldveBeenABridge" or "Queen of No Other Choice". This mean-spirited playfulness, while providing the participants with a memorable experience, can be considered one-sided value co-destruction from the company's point of view, once the incident became widely discussed.

4.5 Antagonism

Another characteristic distinguishing digital value co-creation from that occurring in an unmediated environment is the higher level of antagonism between actors. Since digital interactions between consumers and companies, as well as between consumers themselves, occur in selected communicational hubs, like social media profiles (both official and fan made), dedicated groups, and discussion boards, consumer communities are more likely to surface than in an unmediated environment (when the geographical distance becomes a limiting factor). While members of such communities are usually very passionate about the brand or the product, this usually leads to high expectations as well, which cannot always be met by the company, leading to consumers becoming disillusioned with the company. This leads to a paradox, wherein consumers, despite spending large amounts of time interacting with the company and other members of the community, are highly critical, or even hostile towards the company. At the same time many still remain avid users of the products offered by the said company. Sometimes, these negatively inclined consumers constitute a definite majority of the entire community, or even consumer base. Such a situation can obviously lead to value co-destruction – at least from the perspective of the company. This can be illustrated by the case of Games Workshop, a company that sells tabletop wargaming miniatures. Around this company a particular type of community has been created, one which [7] refer to as a "counter-brand community". Despite loving their games, consumers were frustrated with the high prices of the

miniatures, which led to them turning to 3d printing and other methods of creating substitutes themselves, appropriating the value, which would otherwise go to the company.

However, antagonistic interactions can also lead to value co-creation for both sides. For instance, fans of the Elder Scrolls video game series have been highly critical of its parent company Bethesda. The company became notorious for releasing games lacking features and filled with bugs. However, consumers still enjoy their core experience, and the company provides them with a robust co-creational framework that allows them to create complex modifications (mods) of the game. This leads to a situation where consumers, despite being clearly frustrated with the company, begrudgingly take it upon themselves to fix the bugs and introduce the missing features themselves, and even creating new content for the game. Thanks to this, despite Bethesda's poor reputation amongst consumers and limited support after release, their games have remained popular.

5 Summary of Digital Value Co-creation

Digital value co-creation differs significantly from the co-creation occurring in an unmediated environment. We claim that digital consumer empowerment resulting from digital consumers traits influences the characteristics of digital value co-creation (see Fig. 1).

The traits of modern digital consumers include Internet usage and having access to numerous media channels, information empowerment, accelerated communication (between consumers and companies), interconnectedness, reciprocity between digital and non-digital (unmediated) contexts of own lives, a wider choice of products and services linked with higher expectations and requirements. The resulting characteristics of digital value co-creation are co-presence, automatization, antagonism, playfulness, and simulation.

Interestingly, some of the traits of digital value co-creation seem to pull digital co-creation in different, mutually exclusive directions. Co-presence makes the scope of value co-creation larger, increasing the number of potential co-creational processes taking place, which can leave less time a company can devote to a single interaction. Automatization is a partial answer to this problem; however, these two traits lead to a dehumanisation of the entire process of digital value co-creation.

Conversely, simulation aims to counteract this effect and to reenchant the process of digital value co-creation by projecting personalities onto a brand's social media profiles and making them more human-like. Similarly, playfulness and antagonism also result in reintroducing the human factor into the digital value co-creation, by introducing non-rational and emotional elements into the process—for better or worse.

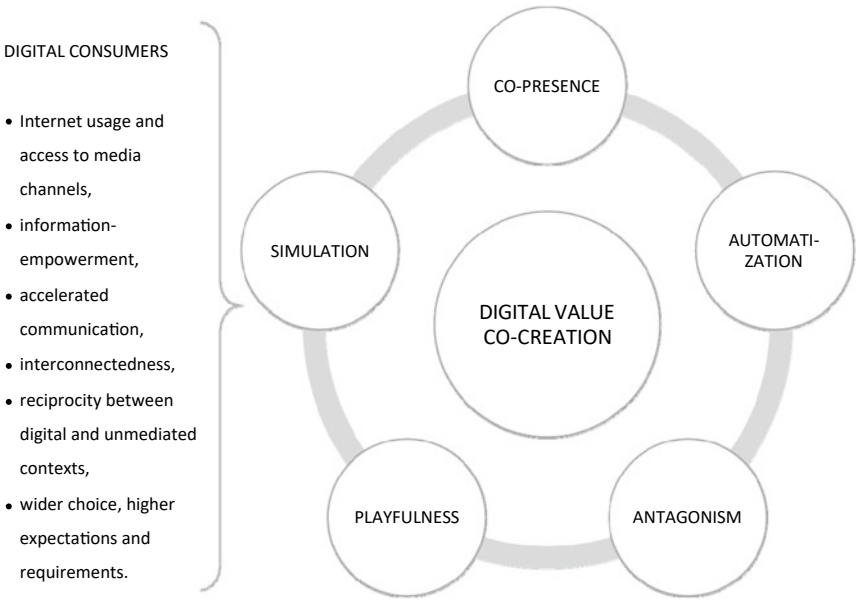


Fig. 1 The characteristics of digital value co-creation resulting from the empowerment of digital consumers

6 Conclusions

The study contributes to the value of co-creation literature, particularly the SDL stream. This is achieved by identifying key characteristics of value co-creation and co- destruction processes that occur in a digital context, as opposed to those that are unmediated by digital technologies. This is particularly important, as digital value co-creation becomes more and more prevalent due to the ever-increasing popularity of social media, mobile devices with access to the Internet, and e-commerce.

From these theoretical developments, some managerial implications can also be derived. First and foremost, business practitioners need to take into account the duality intrinsic to digital consumers. In the digital context, due to co-presence, automatization and simulation, it is possible to increase both the scale and the depth of value co-creation by engaging more consumers in the process and/or to a larger degree. However, due to other characteristics of digital value co-creation, such as playfulness and antagonism, this does not always yield positive results (at least from the perspective of the company), and can oftentimes lead to value co-destruction rather than co-creation. The examples provided in the chapter indicate that despite the theoretical cooperativeness of value co-creation, in the digital context there can also exist a strong competitive element, or even open hostility between different actors. However, this can be somewhat counteracted by properly designing co-creational frameworks and systems, allowing consumers the amount of freedom that reflects the quality of the relationship between the company and its consumers (though it

can of course change dynamically). For instance, should these relationships turn sour, utilising more closed and moderated methods of co-creation (like closed polls) would create fewer opportunities for playful trolling from consumers than more open, unmoderated ones. However, companies boasting an exceptional rapport with their consumers should be able to use more open methods (like hashtags, fan art contests, polls where consumers can add their own propositions) which would allow them to leverage consumers' knowledge, skills and enthusiasm to a higher degree.

As this is a purely theoretical chapter, further research on the subject could involve empirical testing of the prevalence of the specified characteristics in digital and non-mediated value co-creation, as well as its impact on the outcomes of the process. For instance, creative inputs of consumers during online events and real-life fan meetings within the same fandom could be compared to assess the respective playfulness and antagonism levels and establish how they influence the value of these inputs. Similar studies could be carried out for other characteristics as well.

Acknowledgements The project was financed within the Regional Excellence Initiative programme of the Minister of Science and Higher Education of Poland, years 2019-2022, grant no. 004/RID/2018/19, financing 3,000,000.

References

1. Alamäki, A., Korpela, P.: Digital transformation and value-based selling activities: seller and buyer perspectives. Balt. J. Manag. **16**(2), 298–317 (2021). https://doi.org/10.1108/BJM-08-2020-0304
2. Baudrillard, J.: Simulacra and simulation. New York City: Semiotext(e) (1983); Brodie, R.J., Löbler, H., Fehrer, J.A.: Evolution of service-dominant logic: Towards a paradigm and metatheory of the market and value cocreation? Ind. Mark. Manag. **79**(March), 3–12 (2019). https://doi.org/10.1016/j.indmarman.2019.03.003
3. Caridà, A., Edvardsson, B., Colurcio, M.: Conceptualizing resource integration as an embedded process: matching, resourcing and valuing. Mark. Theory **19**(1), 65–84 (2019). https://doi.org/10.1177/1470593118772215
4. Chandler, J.D., Vargo, S.L.: Contextualization and value-in-context: how context frames exchange. Mark. Theory **11**(1), 35–49 (2011). https://doi.org/10.1177/1470593110393713
5. Colm, L., Ordanini, A., Parasuraman, A.: When service customers do not consume in isolation: a typology of customer copresence influence modes (CCIMs). J. Serv. Res. **20**(3), 223–239 (2017). https://doi.org/10.1177/1094670517690025
6. Corsaro, D.: Capturing the broader picture of value co-creation management. Eur. Manag. J. **37**(1), 99–116 (2019). https://doi.org/10.1016/j.emj.2018.07.007
7. Cova, B., White, T.: Counter-brand and alter-brand communities: the impact of Web 2.0 on tribal marketing approaches. J. Mark. Manag. **26**(3–4) (2010). https://doi.org/10.1080/02672570903566276
8. Davenport, T.H., Beck, J.C.: The Attention Economy: Understanding the New Currency of Business. Harvard Business Review Press, Boston (2013)
9. Dey, B.L., Yen, D., Samuel, L.: Digital consumer culture and digital acculturation. Int. J. Inf. Manag. **51**(December 2019) 102057 (2020). https://doi.org/10.1016/j.ijinfomgt.2019.102057
10. Echeverri, P., Skålén, P.: Co-creation and co-destruction: a practice-theory based study of interactive value formation. Mark. Theory **11**(3), 351–373 (2011). https://doi.org/10.1177/1470593111408181

11. Ellegaard, C., Medlin, C.J., Geersbro, J.: Value appropriation in business exchange- literature review and future research opportunities. J. Bus. Ind. Mark. **29**, 185–198 (2014). https://doi.org/10.1108/JBIM-03-2012-0039
12. Firat, A.F., Venkatesh, A.: Postmodernity: the age of marketing. Int. J. Res. Mark. **10**(3), 227–249 (1993). https://doi.org/10.1016/0167-8116(93)90009-N
13. Fırat, A.F., Dholakia, N.: Theoretical and philosophical implications of postmodern debates: some challenges to modern marketing. Mark. Theory **6**(2), 123–162 (2006). https://doi.org/10.1177/1470593106063981
14. Goldhaber, M.H.: The attention economy and the net. First Monday 2(4) (1997). https://doi.org/10.5210/fm.v2i4.519
15. Hulland, J.: Conceptual review papers: revisiting existing research to develop and refine theory. AMS Rev. **10**(1–2), 27–35 (2020). https://doi.org/10.1007/s13162-020-00168-7
16. Manovich, L.: The Language of New Media. MIT Press, Cambridge (2001)
17. Matarazzo, M., Penco, L., Profumo, G., Quaglia, R.: Digital transformation and customer value creation in Made in Italy SMEs: A dynamic capabilities perspective. J. Bus. Res. **123**, 148–2963 (2021). https://doi.org/10.1016/j.jbusres.2020.10.033
18. Morgan-Thomas, A., Dessart, L., Veloutsou, C.: Digital ecosystem and consumer engagement: A socio-technical perspective. J. Bus. Res. **121**(March), 713–723 (2020). https://doi.org/10.1016/j.jbusres.2020.03.042
19. Oxford University Press. (n.d.). Empowerment. In: Oxford Learner's Dictionaries. Retrieved from https://www.oxfordlearnersdictionaries.com/definition/english/empowerment
20. Pathak, B., Ashok, M., Tan, Y.L.: Value co-destruction: Exploring the role of actors' opportunism in the B2B context. Int. J. Inf. Manag. **52**, 102093 (2020). https://doi.org/10.1016/j.ijinfomgt.2020.102093
21. Pera, R., Menozzi, A., Abrate, G., Baima, G.: When cocreation turns into codestruction. J. Bus. Res. **128**(January), 222–232 (2021). https://doi.org/10.1016/j.jbusres.2021.01.058
22. Prahalad, C.K., Ramaswamy, V.: Co-creation experiences: the next practice in value creation. J. Interact. Mark. **18**(3), 5–14 (2004). https://doi.org/10.1002/dir.20015
23. Short, J., Williams, E., Christie, B.: The Social Psychology of Telecommunications. Wiley, London (1976)
24. Simon, H.: Designing Organizations for an Information-Rich World. Johns Hopkins University Press, Baltimore, MD (1971)
25. Srai, J.S., Lorentz, H.: Developing design principles for the digitalisation of purchasing and supply management. J. Purch. Supply Manag. **25**(1), 78–98 (2019). https://doi.org/10.1016/j.pursup.2018.07.001
26. Tkaczyk, J.: Digital consumer: trends and challenges. In: G. Mazurek, J. Tkaczyk (Eds.), The Impact of the Digital World on Management and Marketing, pp. 353–368. Poltext, Warsaw (2016)
27. Vargo, S.L., Lusch, R.F.: Evolving to a new dominant logic for marketing. J. Mark. **68**(1), 1–17 (2004). https://doi.org/10.1509/jmkg.68.1.1.24036;
28. Vargo, S.L., Lusch, R.F.: Institutions and axioms: an extension and update of service-dominant logic. J. Acad. Mark. Sci. **44**(1), 5–23 (2016). https://doi.org/10.1007/s11747-015-0456-3
29. Vargo, S.L., Lusch, R.F.: Service-dominant logic 2025. Int. J. Res. Mark. **34**(1), 46–67 (2017). https://doi.org/10.1016/j.ijresmar.2016.11.001
30. Vargo, S.L., Maglio, P.P., Akaka, M.A.: On value and value co-creation: a service systems and service logic perspective. Eur. Manag. J. **26**(3), 145–152 (2008). https://doi.org/10.1016/j.emj.2008.04.003
31. Venkatesh, A.: Postmodernism perspectives for macromarketing: an inquiry into the global information and sign economy. J. Macromark. **19**(2) (1999). https://doi.org/10.1177/0276146799192006
32. Warner, K.S.R., Wäger, M.: Building dynamic capabilities for digital transformation: an ongoing process of strategic renewal. Long Range Plan. **52**(3), 326–349 (2019). https://doi.org/10.1016/j.lrp.2018.12.001

33. Wieczerzycki, M.: Wymiary nowych mediów a reinterpretacja pojęcia konsumenta w teoriach marketingowych. Marketing i Rynek **10**, 11–18 (2014)
34. Yan, T., Wagner, S.M.: Do what and with whom? Value creation and appropriation in inter-organizational new product development projects Tingting. Int. J. Prod. Econ. **191**, 1–14 (2017). https://doi.org/10.1016/j.ijpe.2017.05.010

Multiclass Prediction of Heart Disease Patients Using Big Data Analytics

Sarita Mishra, Manjusha Pandey, Siddharth Swarup Rautaray, and Sabyasachi Chakraborty

Abstract The rapidly growing rate of illness and death is the result of many diseases. Another major factor is cardiovascular disease (CVD) due to heart failure. According to statistics from around the world, the highest rate of natural death is caused by heart problems. The number of deaths resulting from this can be controlled by the early detection of heart disease chances in a person. Big data and several machine learning technologies have made it possible to discover the chances of a cardiac issue in a person in much advance. Many data scientists have successfully exploited the big data available for heart disease patients and have developed prediction models using different algorithms that are non-invasive, accurate, and appear to be very effective in analyzing patients' characteristics and detecting the presence or absence of heart disease in them. However, to provide appropriate preventive measures and appropriate treatment to patients, it is not enough to detect the presence of CVD, but the degree of impact the disease has left on a person needs to be measured. In this paper, we have compared the performance of five different machine-based algorithms (Logistic Regression, Support Vector Machine, Random Forest, KNN, and Naïve Bayes) which are used to classify the cardiovascular disease into five different classes. 0–4) with the increasing value from 0. These algorithms are used in their most common ways and in the One-vs-All method with the best performance in the latest scenario. The results of this study showed that the KNN algorithm provided 99.56% best predictive accuracy with a combination of One-vs-all and Principal Component Analysis strategies that surpassed all other algorithms.

Keywords Cardiovascular Disease (CVD) · Multiclass Classification · One-vs-All (OVA) · Principal Component Analysis (PCA)

S. Mishra (✉) · M. Pandey · S. S. Rautaray
School of Computer Engineering, KIIT Deemed to Be University, Bhubaneswar, Odisha 751024, India
e-mail: 2181058@kiit.ac.in

S. Chakraborty
Chief Technology Officer, Bhusan, South Korea

S. S. Rautaray et al. (eds.), *Data Science in Societal Applications*, Studies in Big Data 114, https://doi.org/10.1007/978-981-19-5154-1_11

1 Introduction

The human heart is a vital organ about the size of a big fist and weighed 230 to 340 grams. Its role is to oxidize deoxygenated blood and pump it throughout the body as part of the circulatory system. The heart resides in a two-walled sac, called the pericardium, which protects the heart within the chest. The fluid, known as pericardial fluid, flows between two layers of the pericardium that keep the heart lubricated during various heart movements, the diaphragm, and the lungs. The heart performs its function in two cycles: the circulatory system and the circulatory system. In the pulmonary circulation, the oxygen-deprived blood absorbed by the heart during the process of inhalation reaches the lungs through the pulmonary artery, receives oxygen there, and returns to the heart through the pulmonary artery to the left auricle [12, 14]. In systemic circulation, oxygenated blood from the left auricle descends to the left ventricle and eventually leaves the heart through the aorta which separates and divides into many arteries and capillaries that supply oxygenated blood to all parts of the body. Under Fig. 1 shows the workings of the human heart.

Any blockage in any of these blood vessels blocks the smooth flow of blood and may lead to heart attacks. [13, 15] There are many factors that can lead to such blockages including high-cholesterol diets, diet. excessive fat, physical inactivity, stress at work, sleep disturbances, air pollution, alcohol or tobacco use, etc.

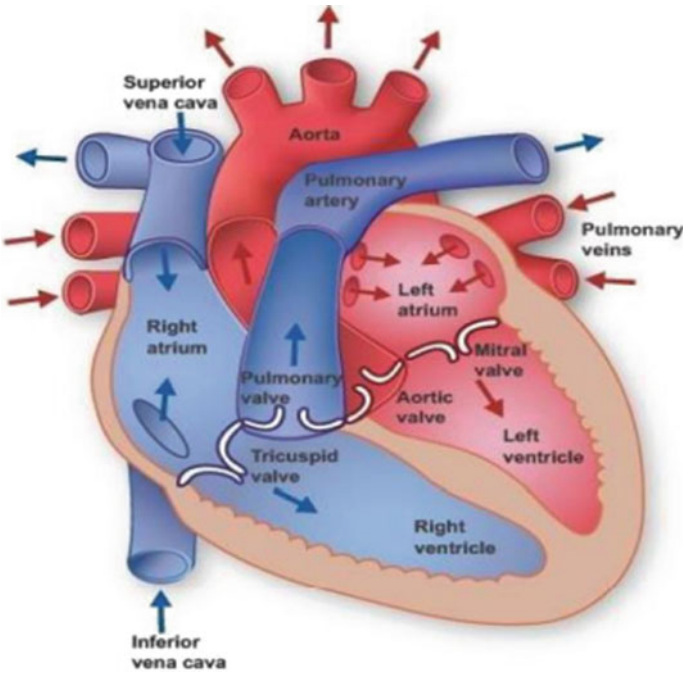


Fig. 1 Working of the human heart

According to the World Health Organization (WHO), approximately 17.9 million people die each year from heart disease [1]. The fast-growing mortality rate can be controlled by predicting the risk of early CVD in a person [2]. Heart disease prognosis can be made with the help of advanced machine-based models that have proven to be very useful for both patients and physicians. Most machine-based predictive models that have been developed so far can differentiate between patients by simply detecting the presence or absence of any heart problem in them. However, in this paper, we have emphasized the division of patients into five categories that reflect the degree of impact of the disease on them and thus provide a deeper understanding of the patient's health status.

In this research work, heart patients are divided into five categories (0 to 4) with a category 0 indicating the absence of the disease, a category 1 showing a minor effect of the disease, and thus an increase in the number of classes showing an increase in disease. criticism of Sect. 4 which means a very critical situation. The algorithms used for this category are Logistic Regression (LR), Vector Support Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Naïve Bayes (NB). Here we have compared the performance of these algorithms with their standard methods and their use of One-vs-All (OVA) in which they create a multi-stage multi-level division instead of a single multi-phase division. The results of this research work showed that the latter provides much better accuracy than the first. OVA-based performance is further enhanced by the Key Component Analysis (PCA) system which has increased the predictive accuracy of algorithms with KNN showing the best performance with 99.56% accuracy.

This paper is also organized as follows. The next Sect. 2 discusses a few previous activities regarding heart disease prediction using machine learning models. Section 3 presents the workflow of this study, the data used, the preliminary processing of the data, and the algorithms used. Section 4 discusses the results obtained during this study and the paper ends up in Sect. 5.

2 Literature Review

As mentioned in the previous section, most of the earlier models for heart disease were designed for binary segregation of patients, so our state-of-the-art base has been reduced to a limited number of research activities for many cardiovascular categories. a disease in the health care sector with few binary separation functions.

Kirsi Varpa et al. [3] performed multiple classifications in Otoneurological disorder patients using KNN and SVM classification algorithms. These algorithms are used in their conventional methods as well as using the OVA method where the latter gives the best results with KNN. Anurag Kumar Verma et al. [4], in their paper, classified people suffering from skin diseases into six distinct categories (psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra). This classification is done using six machine learning algorithms and producing their own ensembles using bagging, Adaboost, and gradient boosting. It

was noted that the use of ensembles to produce predictive models provides better results than each algorithm.

Hin Wai Lui et al. [5] merged the normal neural network with the convolutional neural network to form a multiclass classification of patients affected by Myocardial Infarction disease. The upgraded model was able to provide 97.2% accuracy with 92.4% sensitivity and can be plugged into portable devices. C. Beulah Christalin Latha et al. [6] performed two classifications of cardiovascular patients in the Cleveland heart disease database. They used six different algorithms: Bayes Net, Naïve Bayes, Random Forest, C4.5, Multilayer perceptron, and Projective Adaptive Resonance Theory (PART) to create their ensembles again. The combining methods used were bagging, voting, and packaging. It has been noted that the use of manufactured merged models offers much higher accuracy than individual weak algorithms, with a multi-voting system showing a higher accuracy with an increase of 7%. This accuracy is also enhanced by the use of the PCA process.

Abderrahmane El.daoudy et al. [1] created a cardiovascular prediction model using Apache Cassandra to store highly generated data and Spark MLlib to make predictions. This model, built using a random forest algorithm was able to handle real-time data and provided 87.5% accuracy and 86.67% sensitivity.

3 Method and Materials

This section describes the flow of work during this research, the dataset used, its analysis and preprocessing, and also explains the various algorithms used.

The steps presented in the flowchart in Fig. 2 are performed for the One-vs-all implementation of each of the machine learning algorithms used, which are, logistic regression, SVM, random forest, KNN, and Naïve Bayes such that the binary classifiers generated from each algorithm give their own accuracies whose mean is estimated to determine the accuracy of that algorithm.

3.1 Dataset Description

The dataset used for this research work is the Cleveland heart disease dataset available in the UCI machine learning repository. This dataset has the target variable which is a multivariate attribute and can take up values between 0 and 4 with class 0 indicating an absence of the disease, class 1 indicating a mild impact of the disease, class 2 indicating a moderate state, class 3 implies a slightly severe state, and class 4 indicates the most critical state.

The other predictor variables that are involved in the classification process are the patient's age, gender, type of chest pain, blood pressure while resting, cholesterol level in the blood, blood sugar level, maximum heartbeat rate, and a few others [7]. The 13 predictors along with the target variable are presented in Table 1.

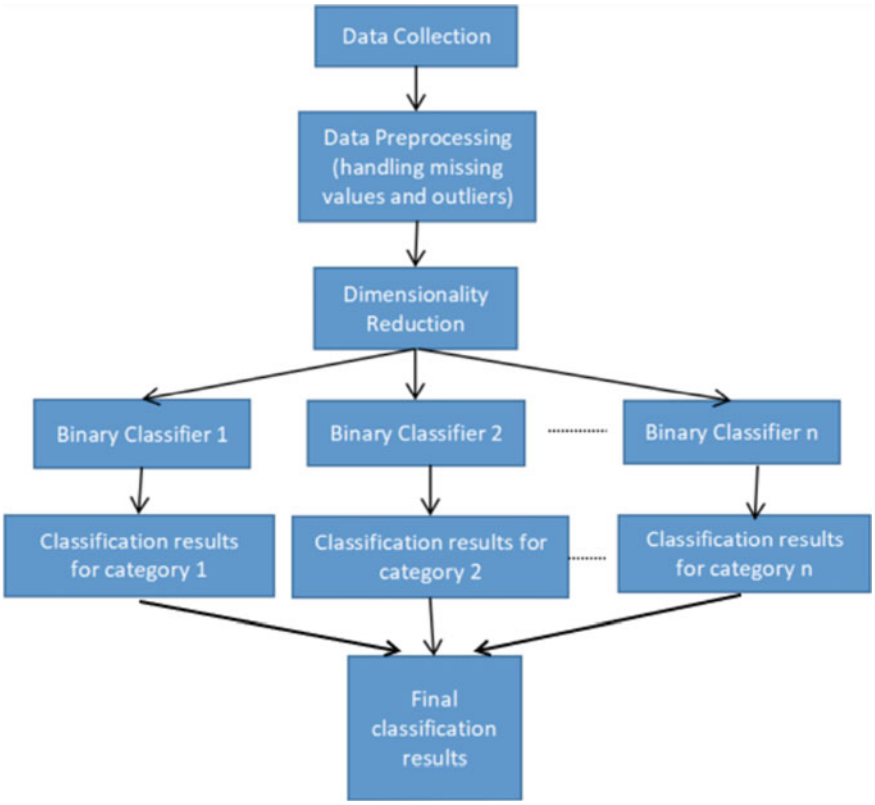


Fig. 2 Flow of work

Before applying the various machine learning algorithms to predict the target attribute values, the dataset needs to be preprocessed and analyzed properly to handle the outliers and missing values, and to identify the relevant patterns hidden in it.

In the dataset used, the Ca and Thal attributes contained missing values that we replaced with the values 0 and 3 which have the highest count in these attributes, respectively. The outliers are detected by plotting the boxplots of the attributes and it was observed that their presence did not affect the prediction accuracy adversely, hence remained undisturbed.

The following Fig. 3 shows the bar plot for the target attribute, Figs. 4 and 5 show the boxplot representation of the age and chol attributes, respectively.

It can be observed from Fig. 3 that the maximum number of people in the dataset (more than 150) belong to class 0, i.e., they are completely free from CVD and class 4 contains the minimum number of people. Figure 4 indicates that the people belonging to a higher age group are more prone to being in class 4, i.e., they are more likely to reach a critical state with a few outliers where people with an age less than 40 are in a critical state, i.e., they belong to class 4. Figure 5 indicates that the average

Table 1 Features of the Cleveland heart disease dataset

Feature	Description	Range
Age	Age of the patient	29–70
Sex	Gender of the patient	0-Female 1-Male
Cp	Type of chest pain	1-Typical Angina 2-Atypical Angina 3-Non-angina 4-Asymptomatic
Trestbps	Blood pressure of the patient while resting in mm Hg	94–200
Chol	Serum cholesterol in mg/dl	126–564
Fbs	Blood sugar level due to fasting in mg/dl	0, 1
Restecg	ECG result while resting	0, 1, 2
Thalach	Maximum rate of heartbeat	71–202
Exang	Angina induced while exercising	0, 1
Oldpeak	Relative ST depression while exercising and resting	1, 2, 3
Slope	Slope of peak ST segment	1, 2, 3
Ca	Number of major blood vessels that are colored by fluoroscopy	0, 1, 2, 3
Thal	Type of defect	3-Normal defect 6-Fixed defect 7-Reversible defect
Target	Level of impact of the disease on the patient/target class of the patient	0-absences of CVD 1-mild impact of CVD 2-moderate impact of CVD 3-slightly severe impact of CVD 4-highly critical state

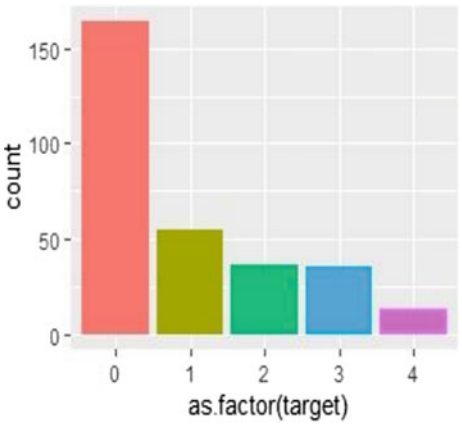


Fig. 3 Bar plot for target variable

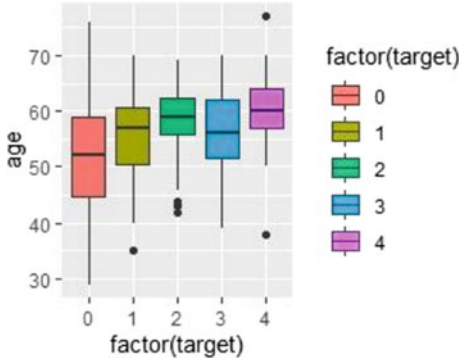


Fig. 4 Boxplot for age attribute

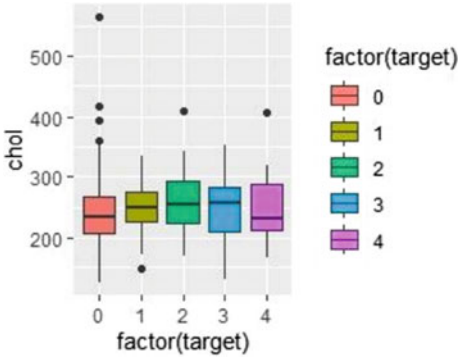


Fig. 5 Boxplot for cholesterol attribute

cholesterol of all the people falls in the range of 200–300. Also, it can be seen from Fig. 5 that few people having a cholesterol level close to 500 are completely free from heart disease.

Figure 6 represents the correlation matrix for the dataset used where every value less than 0 indicates a negative correlation, every value greater than 0 indicates a positive correlation and a 0 value indicates complete independence between the two associated attributes.

3.2 Algorithms Used

In this research work, we have used the following five algorithms to classify the patients and compared their performances with and without the One-vs-All approach.

Logistic Regression, a supervised classification algorithm, is typically used for binary classification and cannot perform multiclass classification in its ordinary form; however, in this research, we have used it for multiclass classification by implementing it with the One-vs-all approach. LR method can be used on datasets that are

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1	-0.1	0.1	0.28	0.21	0.12	0.15	-0.39	0.09	0.2	0.16	0.37	0.13	0.22
sex	-0.1	1	0.01	-0.06	-0.2	0.05	0.02	-0.05	0.15	0.1	0.04	0.09	0.38	0.22
cp	0.1	0.01	1	-0.04	0.07	-0.04	0.07	-0.33	0.38	0.2	0.15	0.23	0.26	0.41
trestbps	0.28	-0.06	-0.04	1	0.13	0.18	0.15	-0.05	0.06	0.19	0.12	0.1	0.13	0.16
chol	0.21	-0.2	0.07	0.13	1	0.01	0.17	0	0.06	0.05	0	0.12	0.02	0.07
fbs	0.12	0.05	-0.04	0.18	0.01	1	0.07	-0.01	0.03	0.01	0.06	0.14	0.06	0.06
restecg	0.15	0.02	0.07	0.15	0.17	0.07	1	-0.08	0.08	0.11	0.13	0.13	0.02	0.18
thalach	-0.39	-0.05	-0.33	-0.05	0	-0.01	-0.08	1	-0.38	-0.34	-0.39	-0.27	-0.27	-0.42
exang	0.09	0.15	0.38	0.06	0.06	0.03	0.08	-0.38	1	0.29	0.26	0.15	0.33	0.4
oldpeak	0.2	0.1	0.2	0.19	0.05	0.01	0.11	-0.34	0.29	1	0.58	0.3	0.34	0.5
slope	0.16	0.04	0.15	0.12	0	0.06	0.13	-0.39	0.26	0.58	1	0.11	0.29	0.38
ca	0.37	0.09	0.23	0.1	0.12	0.14	0.13	-0.27	0.15	0.3	0.11	1	0.26	0.52
thal	0.13	0.38	0.26	0.13	0.02	0.06	0.02	-0.27	0.33	0.34	0.29	0.26	1	0.51
target	0.22	0.22	0.41	0.16	0.07	0.06	0.18	-0.42	0.4	0.5	0.38	0.52	0.51	1

Fig. 6 Correlation matrix for cleveland heart disease dataset attributes

free from missing values. But the dataset used here has missing values in Ca and Thal attributes which we have handled by replacing the blanks with the value having the highest frequency in the respective attribute. The core functionality of this algorithm that is used to estimate the probability of a specific class being applicable on a data point is the sigmoidal function as shown in Eq. 1 [8].

$$y = 1/\left(1 + e^{-x}\right) \tag{1}$$

where 'e' is the Euler constant with the value 2.718, x is the linear combination of all the predictors, and y is the probability between 0 and 1 indicating the class to which the new tuple belongs by considering a threshold value between 0 and 1, which is 0.5 by default.

Support Vector Machine is another supervised algorithm, that is applicable for binary as well as multiclass classification. It attempts to generate a separating boundary, known as a hyperplane, depending on the dataset being used that separates the tuples belonging to vivid classes with the maximum margin [9]. Out of all the possible hyperplanes that segregate the data points, the hyperplane that provides the maximum margin is called the Maximal Margin Hyperplane (MMH). Figure 7

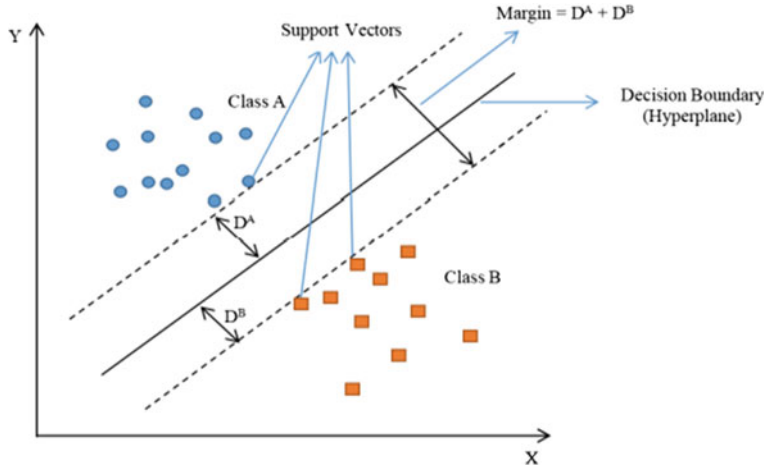


Fig. 7 Support vector machine architecture in a 2-d space

shows a 2-dimensional space containing data points of two different classes A and B separated by a hyperplane.

In the case of non-linearly separable data, that is when the hyperplane required to separate the data points is not a straight line, a kernel function is used that takes as input the low-dimensional feature space and converts it into a high-dimensional feature space and generates a curve or a plane as a hyperplane to separate the data elements belonging to distinct classes. While implementing SVM for our dataset, without the One-vs-all approach, 182 support vectors were obtained as shown in Fig. 8.

```
> hsvm <- svm(target ~ .,data=traindsvm,cost=85,gamma = 0.1)
> summary(hsvm)

Call:
svm(formula = target ~ ., data = traindsvm, cost = 85, gamma = 0.1)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  85
      gamma:  0.1
    epsilon:  0.1


Number of Support Vectors:  182
```

Fig. 8 SVM model built

Random Forest is a machine learning classifier based on the ensemble technique and uses the concept of decision trees in a randomized manner [10]. Each decision tree generated takes up its own set of tuples and attributes, known as a bootstrap, from the original dataset and comes up with its prediction result for the newly introduced data point. These individual prediction outputs are aggregated using the majority voting technique, i.e., the prediction value that is generated by the maximum number of trees as their output is considered to be the final prediction result of the random forest model [11].

Nearest Neighbors is another supervised machine learning algorithm used to solve classification and regression problems. To classify a newly introduced data point into one of the classes in the dataset, 'K' data points that are closest to the new data point are identified. These 'K' nearest data points are detected by measuring the distance of all the existing data points from the new data point. The class or category to which majority of these 'K' neighbors belong is the desired class to which the new data point should be assigned. The distance between the new data point and existing data points can be calculated using several measures [10]. Few commonly used techniques are given as follows.

Euclidean distance given by

$$d = \left(\sum \left(x_i - y_i\right)^2\right)^{1/2} \tag{2}$$

Manhattan distance given by

$$d = \sum \left(x_i - y_i\right) \tag{3}$$

Minkowski distance given by

$$d = \left(\sum \left(x_i - y_i\right) s^r\right)^{1/r} \tag{4}$$

The error in prediction varies with the value chosen for the variable 'K'. In our research, we iterated over 1 to 10 to choose the value of 'K' which will give the minimum error rate. The following Figs. 9, 10 and 11 show the error rate versus K plots for the KNN algorithm implemented in its ordinary form, with One-vs-all, and with OVA combined with the principal component analysis technique.

It can be observed from the above figures that the error rate is minimum for $K = 10$ for ordinary KNN implementation, reaches a minimum at $K = 2$, and remains same till $K = 10$ when implemented with OVA, and the error rate becomes 0 for $K = 1$ to 10 when KNN is implemented with OVA combined with PCA.

Naïve Bayes is also a supervised learning algorithm that is also a probabilistic classifier. This algorithm assumes complete independence among the features of the dataset, i.e., occurrence of one attribute does not depend on the occurrence of any other attribute [10]. It works on the principle of the Bayes theorem which is given as follows:

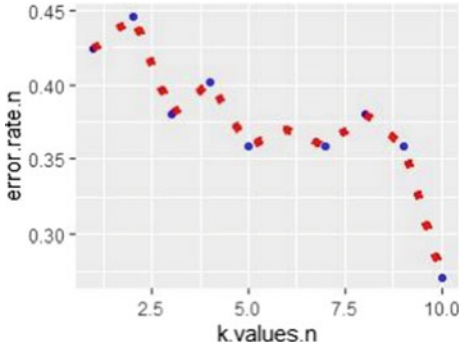


Fig. 9 Ordinary KNN

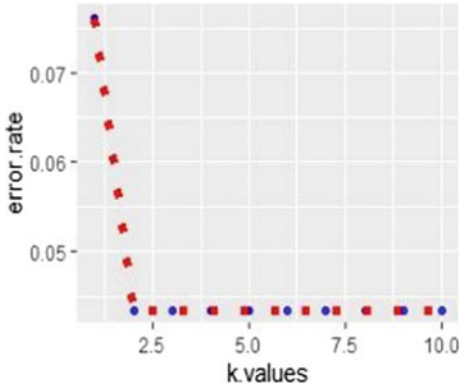


Fig. 10 One-vs-all

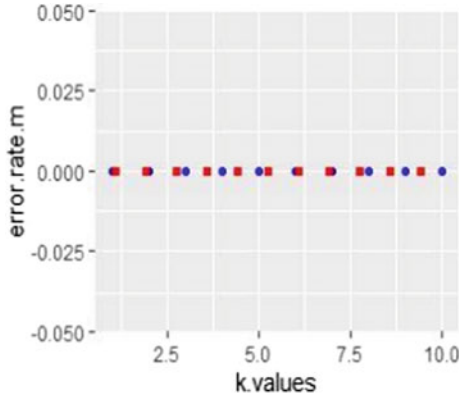


Fig. 11 One-vs-all & PCA

$$P(A|B) = (P(B|A) * P(A))/P(B) \quad (5)$$

where P(A|B) is the probability of hypothesis A when event B has occurred, P(B|A) is the likelihood of occurrence of an event given that A is true, P(A) denotes the prior probability of the hypothesis before the event has occurred, P(B) denotes the evidence, i.e., probability of occurrence of event B.

One-vs-All is a technique used to implement machine learning algorithms to perform multiclass classification much more efficiently compared to their performance without it. In this approach, 'n' binary classifiers are built based on a chosen algorithm instead of a single multiclass classifier, where 'n' is the number of classes in the dataset. Each binary classifier is committed toward a single class, i.e., each binary classifier gives the accuracy in prediction for a single class by considering that class as class 1 and all other classes as class 0. The accuracies obtained in making predictions for each class as provided by their associated binary classifiers are averaged to find the overall accuracy of that algorithm. Figure 12 shows the architecture of the One-vs-all technique.

Principal Component Analysis is an unsupervised machine learning algorithm that is used to reduce the dimensionality of the dataset thus allowing the model to predict the target variable values for a reduced dataset thus reducing the chances of overfitting [10]. PCA tries to find attributes, known as Principal Components, that provide maximum variance in the higher dimensional space and project those onto a smaller dimensional space retaining only the relevant information. Using these principal components to make predictions not only reduces the burden on the models in making predictions but also improves the accuracy of the prediction by the model.

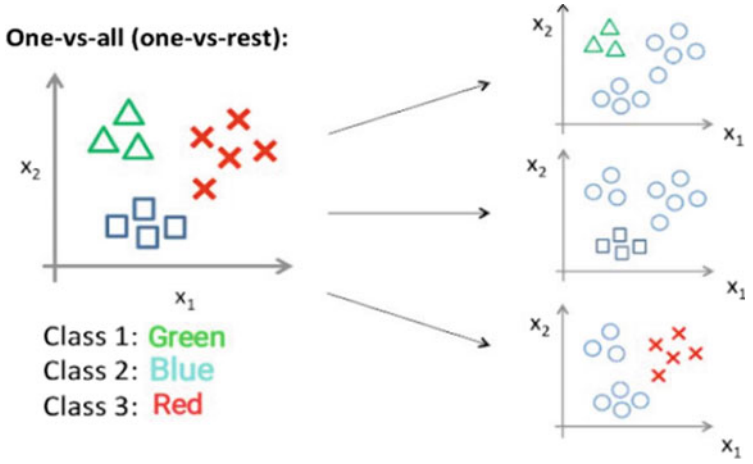


Fig. 12 Schematic diagram of one-vs-all technique

4 Experimental Results

The research outcome shows a comparison among the heart disease prediction accuracies provided by the five algorithms: LR, SVM, RF, KNN, and NB when implemented on the Cleveland heart disease dataset. These algorithms are implemented in three ways: in their ordinary form, with the One-vs-all approach, and One-vs-all and principal component analysis combined. The accuracies obtained in each case are compared and it is observed that using the OVA approach significantly increases the accuracy in prediction for each algorithm. Also, using the OVA technique on the dataset reduced by PCA further enhances the performance of the models.

The accuracy of each algorithm is computed using the confusion matrices generated. Confusion matrix refers to an nxn matrix where n is the number of distinct classes available in the dataset. In a confusion matrix, each column sum indicates the number of people that actually belong to that class and each row sum indicates the number of people who have been categorized into that class. Hence, it can be concluded that the diagonal elements of the confusion matrix indicate the number of patients who have been correctly classified. The Eq. 6 is used to compute the accuracy in classification from the confusion matrix generated.

$$\text{Accuracy} = \left\{ \sum(\text{diagonal elements}) / \sum(\text{all elements}) \right\} * 100 \tag{6}$$

The following Figs. 13 and 14 depict the generated confusion matrices by all the above-mentioned algorithms without OVA and with OVA, respectively. In the OVA approach, as already mentioned earlier, instead of a single multiclass classifier, 5 binary classifiers are built for each algorithm that generates their own 2x2 confusion matrices. The accuracies of all 5 confusion matrices are averaged to compute the overall accuracy provided by that algorithm when implemented with the OVA approach.

The confusion matrices generated by the application of OVA approach on the dataset reduced by PCA have slightly better values than the confusion matrices without PCA, thus providing slightly better accuracy.

The following Table 2 states the accuracy provided by each of the algorithms that are implemented without OVA method, with the OVA approach, and OVA on the PCA reduced dataset.

Figure 15 shows the bar graph representation of the accuracy values acquired by all the algorithms for the three types of implementation. It can be observed from Table 2 and Fig. 15 that no accuracy value exists for the ordinary implementation of logistic regression as it is a binary classification algorithm and can perform multiclass classification only with the One-vs-all approach.

Also, it can be seen that the KNN algorithm has outperformed all other algorithms by providing the highest accuracy of 99.56%.

SVM	0	1	2	3	4
0	45	3	0	2	0
1	2	9	2	3	1
2	2	2	6	2	1
3	0	3	2	3	1
4	0	0	1	1	1

RF	0	1	2	3	4
0	41	3	0	0	0
1	6	11	3	5	0
2	2	3	8	3	3
3	0	0	0	3	1
4	0	0	0	0	0

KNN	0	1	2	3	4
0	45	6	5	2	0
1	4	6	3	5	1
2	0	4	1	1	1
3	0	1	1	2	2
4	0	0	1	1	0

NB	0	1	2	3	4
0	45	4	2	0	0
1	2	8	2	2	1
2	0	1	3	3	1
3	1	4	4	6	0
4	1	0	0	0	2

Fig. 13 Confusion matrices without one-vs-all

5 Conclusion

This paper has emphasized the classification of CVD patients into more than two classes that will be more helpful for the physicians in providing the best possible treatment to their patients with more precision instead of simply discovering the sign of any cardiac problem in them. We have recommended the exploitation of this One-vs-all method with different machine learning algorithms to perform a multiclass classification of the patients. The obtained accuracy is further improved by implementing these algorithms with OVA on the dataset reduced by PCA. Out of the five algorithms implemented during this research, KNN has shown the best performance with an accuracy of 99.56% with a combination of PCA and One-vs-All techniques.

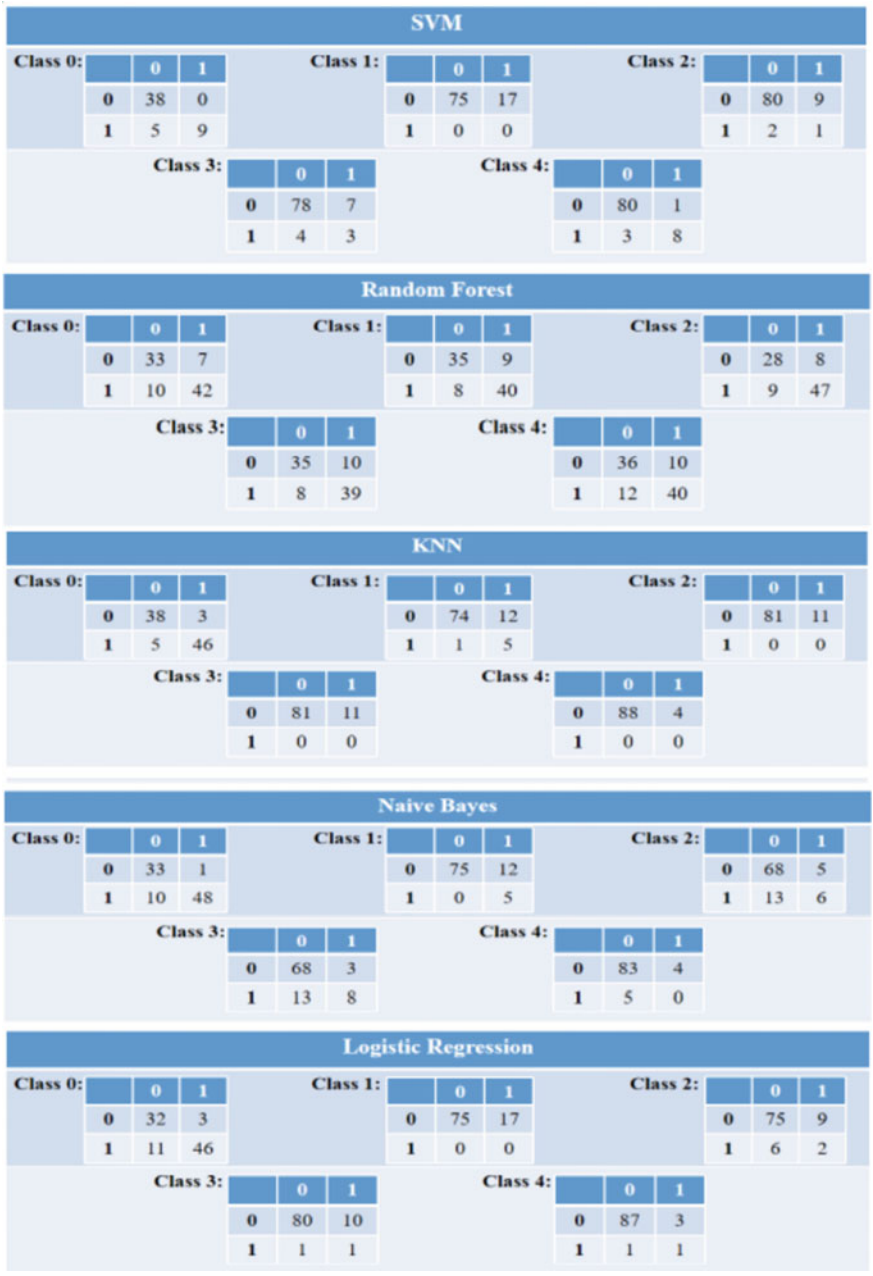


Fig. 14 Confusion matrices with one-vs-All

Table 2 Accuracies obtained by each algorithm in all 3 cases

Algorithm	Accuracy without OVA (%)	Accuracy with OVA (%)	Accuracy with OVA and PCA (%)
SVM	69.57	89.57	89.57
Random forest	68.48	89.13	92.53
KNN	7.83	89.78	99.56
Naïve Bayes	69.57	85.65	85.71
Logistic		86.74	93.41

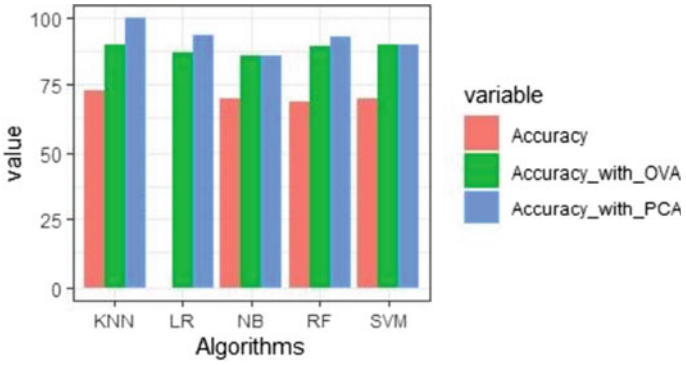


Fig. 15 Bar plot for accuracies obtained

References

1. Ed-Daoudi A, Maalmi, K.: Real-time machine learning for early detection of heart disease using big data approach. In: 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco, 2019, pp. 1–5. https://doi.org/10.1109/WITS.2019.8723839
2. Baitharu, T.R., Pani, S.K: Analysis of data mining techniques for healthcare decision support system using liver disorder dataset. In: International Conference On Computational Modelling And Security (CMS 2016). Procedia Compuer Science 85(2016), pp. 862–870
3. Varpa, K., Joutsijoki, H., Iltanen, K., Juhola, M.: Applying one-vs-one and one-vs-all classifiers in k-nearest neighbors method and support vector machines to an otoneurological multiclass problem. In: Article in Studies in Health Technology and Informatics-January 2011. 10.3233|978-1-60750-806-9-579
4. Verma, A.K., Pal, S., Kumar, S.: Comparison of Skin Disease Prediction by Feature Selection Using Ensemble Data Mining Techniques. Informatics In Medicine Unlocked, vol. 16, p. 100202 (2019). 10/1016/j.imu.2019.1002.02
5. Lui, H.W., Chow, K.L.: Multiclass classification of myocardial infarction with convolutional and recurrent neural networks for portable ECG devices. Inform. Med. Unlocked **13**, 26–33 (2018)
6. Latha, C.B.C., Jeeva, S.C.: Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inform. Med. Unlocked **16**, 100203. https://doi.org/10.1016/j.imu.2019.100203

7. Saxena, K., Sharma, R.: Efficient heart disease prediction system. Procedia Comput. Sci. **85**, 962–969 (2016). https://doi.org/10.1016/j.procs.2016.05.288
8. Bagley, S.C., White, H., Golomb, B.A.: Logistic regression in the medical literature::standards for use and reporting with particular attention to medical domain **54**(10), 979–985 (2001). https://doi.org/10.1016/s0895-4356(01)00372-9
9. Mishra, S., Pandey, M., Rautaray, S.S., Gourisaria, M.K.: A survey on big data analytical tools & techniques in healthcare sector. Int. J. Emerg. Technol. **11**(3), 554–560
10. Lebedev, A.V., Wesman, E., Van Westen, G.J.P.: Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness **6**, 115-125 (2014). https://doi.org/10.1016/j.nicl.2014.08.023
11. Kamran, M., Javed, A.: A survey of recommender systems and their application in healthcare. In: Technical Journal, University of Engineering and Technology (UET) Taxila, Pakistan, vol. 20 No. IV-2015
12. Alarsan, F.I., Younes, M.: Analysis and classifcation of heart diseases using heartbeat features and machine learning algorithms. Alarsan and Younes J Big Data **6**, 81 (2019). https://doi.org/10.1186/s40537-019-0244-x
13. Salma Banu, N.K, Swamy, S.: Prediction of heart disease at early stage using data mining and big data analytics: a survey. In: 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT). 978-1-5090-4697-3/16/
14. Saboji, R.G., Ramesh, P.K.: A scalable solution for heart disease prediction using classification mining technique. In: International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017). 978-1-5386-1887-5/17/
15. Karthick, D., Priyadharshini, B.: Predicting the chances of occurrence of Cardio Vascular Disease (CVD) in people using Classification Techniques within fifty years of age. In: Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018) IEEE Xplore Compliant - Part Number:CFP18J06-ART. ISBN:978-1-5386-0807-4; DVD Part Number:CFP18J06DVD, ISBN:978-1-5386-0806-7