AIHealth 2024

The First International Conference on AI-Health

ISBN: 978-1-68558-136-7

March 10th –14th, 2024

Athens, Greece

AIHealth 2024 Editors

Maura Mengoni, Polytechnic University of Marche, Italy

Amina Souag, Canterbury Christ Church University, UK

AIHealth 2024

Foreword

The First International Conference on AI-Health (AIHealth 2024), held between March 10 – 14, 2024, covered topics blending Artificial Intelligence and health sciences.

Quality healthcare should be extended to all communities. Independent of how big and complex the healthcare systems are, physicians are under increasing time and workload pressures and spending less time with patients. The challenge to deliver high-quality healthcare against administrative burdens is big and increasing.

Healthcare facilities also produce great amounts of data and record high volumes of patient records information. This information is valuable and necessary to quality patient care. This information requires an enormousness effort (time, personnel) to be timely processed for prediction, evaluation and monitoring patients' health.

Artificial Intelligence (AI) comes to rescue in terms of accuracy, precision, rapidity and processing a large volume of data. AI-based health systems benefit for recent advances in sophisticated AI mechanisms for predicting patient health conditions (personalized, at large scale), producing useful analytics on varii patient health aspects, as well as monitoring and controlling patient under scrutiny.

We take here the opportunity to warmly thank all the members of the AIHealth 2024 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to AIHealth 2024.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the AIHealth 2024 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that AIHealth 2024 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the fields of AI and health sciences.

We are convinced that the participants found the event useful and communications very open. We also hope that Athens provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city

AIHealth 2024 Chairs:

AIHealth 2024 Steering Committee
Les Sztandera, Thomas Jefferson University, USA
Hesham H. Ali, University of Nebraska at Omaha, USA
Maura Mengoni, Università Politecnica delle Marche, Italy
Vitaly Herasevich, Mayo Clinic, USA

AIHealth 2024

Committee

AIHealth 2024 Steering Committee
Les Sztandera, Thomas Jefferson University, USA
Hesham H. Ali, University of Nebraska at Omaha, USA
Maura Mengoni, Università Politecnica delle Marche, Italy
Vitaly Herasevich, Mayo Clinic, USA

AIHealth 2024 Technical Program Committee

Hesham H. Ali, University of Nebraska at Omaha, USA
Alireza Atashi, Tehran University of Medical Sciences, Iran
Michael Beigl, Karlsruhe Institute of Technology (KIT), Germany
Sid-Ahmed Berrani, Ecole Nationale Polytechnique, Algiers, Algeria
Elizabeth Borycki, University of Victoria, Canada
An Braeken, Vrije Universiteit Brussel, Belgium
Philippe Cinquin, CHUGrenoble Alpes, France
Marcos Cordeiro d'Ornellas, Universidade Federal de Santa Maria (UFSM) | Hospital Universitário (HUSM), Brazil
Manuel Domínguez-Morales, University of Sevilla, Spain
Sai Anvesh Durvasula, Parabole.ai, USA
Duarte Duque, 2Ai - School of Technology | IPCA, Portugal
Vitaly Herasevich, Mayo Clinic, USA
Haralampos Karanikas, University of Thessaly, Greece
Sushil K. Meher, All India Institute of Medical Sciences, New Delhi, India
Maura Mengoni, Polytechnic University of Marche, Italy
Daniela Micucci, University of Milano - Bicocca, Italy
George Mihalas, Victor Babes Univ. Med.&Pharm, Timisoara | Academy of Medical Sciences, Com. Medical Informatics & Data Protection, Romania
Kartik Palani, iManage / University of Illinois Urbana-Champaign, USA
Nadav Rappoport, Ben-Gurion University of the Negev, Israel
Stefano Rinaldi, University of Brescia, Italy
Floriano Scioscia, Polytechnic University of Bari, Italy
Gro-Hilde Severinsen, Norwegian Center for e-health research, Norway
Jaideep Srivastava, University of Minnesota, USA
Dalibor Stanimirovic, University of Ljubljana, Slovenia
Les Sztandera Thomas Jefferson University, USA
Hamid Usefi, Memorial University, Canada

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

The Role of Artificial Intelligence and Machine Learning in Predictive Health Care, Diagnostics, and Personalized Treatment for Seniors

Rüdiger Höfert
Chief Executive Officer
Absolute Software
Hamburg, Germany
ruediger.hoefert@absolute.de

Wolfgang Bench
Business Development Manager
Absolute Software
Hamburg, Germany
wolfgang.bench@absolute.de

***Abstract* - This paper introduces an AI-based assistance system for elderly care, directly aligning with the conference themes "AI-based Health systems and applications," "Personalized health devices and mobile services," and "Assisted-living applications using affective computing." Addressing elderly loneliness and staff shortages, it leverages data analysis, pattern recognition, and big data analytics with AI frameworks such as TensorFlow and Keras to enhance seniors' quality of life through interactive communication, personalized dialogues, cognitive stimulation, and emergency responses. Our approach, an AI avatar family on tablets, emphasizes empathy over traditional solutions like robotics and wearables, and aims to mitigate loneliness—a key factor in health deterioration. Unlike asserting superiority, this project explores empathy's potential to complement existing technologies, with our exploratory efforts showing promise in early evaluations. By prioritizing empathetic connections and evaluating its impact on user satisfaction and quality of life, our work offers a hopeful, yet cautious, perspective on improving elderly care. This paper concludes that a scalable, empathetic solution is well suited for dealing with the described challenges, and offers a meaningful alternative to existing solutions.**

Keywords-Artificial Intelligence; Elderly Care; Avatar; Machine Learning; Empathy

I. INTRODUCTION

The health care provision for seniors faces a challenge that extends beyond immediate medical care: loneliness. More than one in five seniors suffers from the psychosocial and physical consequences of social isolation, which often leads to a downward cascade in functional competence. The situation is further exacerbated by a shortage of skilled workers in the nursing sector. 770 million people are aged 65 or older. With the total number of individuals requiring care surpassing a billion, the caregiver to care recipient ratio is expected to decrease from 7:1 in 2011 to 3:1 by 2050. These staffing gaps inevitably lead to a deterioration in the quality of care, which directly affects the health and well-being of seniors. With aging, cognitive, social, and physical abilities diminish, leading to increased loneliness and anxiety in the elderly. Families and institutions face limitations in providing adequate help. Emergencies often go unnoticed, with alarm buttons frequently out of reach when needed. Limited interactions and challenges accelerate the speed of physical and mental decline [1].

Artificial Intelligence (AI) and Machine Learning (ML) have the potential to improve the quality of care and relieve the staff [2]. These technologies promise transformative changes in predictive health care, diagnostics, and personalized treatment, especially for the older population [3]. This paper addresses the question of how AI and ML technologies can be applied in elderly care to meet these challenges. It discusses how AI-supported decision-support and health monitoring systems can contribute to improving the quality of care and relieving the burden on nursing staff, without replacing the human component [4].

The primary research question targeted with our solution for elderly care is how the factor of empathy alleviates loneliness and thereby also plays a significant role in stopping cognitive decline and illnesses that are typically accelerated by isolation. The purpose of this paper is to provide a general overview of the benefits of AI in elderly care and to outline the solution as such, with its characteristics, goals, functionality, challenges, and outlook.

The limitations of our approach are rooted in the fact that we have just started the implementation in care facilities. While initial feedback from patients and staff is good, a thorough analysis is still pending. This uncertainty underscores the exploratory nature of our current phase, emphasizing the potential for future developments and refinements based on comprehensive evaluations.

This paper is structured to guide the reader through our comprehensive approach to integrating AI in elderly care as follows:

State of the Art Solutions: This section reviews current elderly care technologies, highlighting how AI avatars offer a unique solution to the limitations of existing applications, especially in addressing loneliness.

Methods, Materials, and Tools: We discuss the technological underpinnings of our AI solution,

 ISBN: 978-1-68558-136-7

emphasizing the selection of AI frameworks and tools that prioritize adaptability, scalability, and empathetic engagement.

Description of the Solution: The AI solution's empathetic design is elaborated, showing how it aims to mitigate elderly loneliness through personalized and compassionate interactions.

Further Aspects and Implementation: Challenges in integrating the solution into care facilities, user acceptance strategies, and privacy considerations are outlined.

Results and Experiences: Initial feedback from the deployment of our AI system in care facilities is shared, highlighting its impact on care quality and areas for improvement.

Conclusion and Future Work: The paper ends with a discussion on the future potential of AI in elderly care, considering scalability and upcoming advancements, and situates our work within the broader AI and healthcare narrative.

II. STATE OF THE ART SOLUTIONS

Elderly care technology has seen significant advancements aimed at improving the quality of life for seniors. These innovations typically fall into several categories, each designed to address specific aspects of elderly care.

A. Robotics

Robotic solutions in elderly care, such as companion robots and robotic assistants, have been developed to provide physical assistance, social interaction, and monitoring. Examples include robots that help with mobility, perform simple household tasks, or offer reminders for medication.

B. Smart Applications

Smart applications and devices, ranging from wearable health monitors to smart home systems, are widely used to enhance safety and health monitoring. These applications can track vital signs, detect falls, and enable remote communication with caregivers and medical professionals.

C. Virtual and Augmented Reality

Virtual Reality (VR) and Augmented Reality (AR): VR and AR technologies are emerging as tools for cognitive stimulation and social interaction. They offer immersive experiences that can help seniors engage in virtual travel, memory exercises, and social activities, potentially reducing feelings of isolation.

D. Benefits and Limitations of the Existing

While these solutions perform valuable tasks and address daily challenges for the elderly, such as improving physical health, ensuring safety, and providing some level of social interaction, they often do not fully address the deeper, long-term problem of isolation and loneliness. This gap is where our proposal, an AI Avatar family on tablets focusing on empathy, introduces a novel approach. Our solution is designed to go beyond the functionality of current technologies by prioritizing emotional intelligence and empathetic interaction. Recognizing that loneliness and social isolation have profound effects on the mental and physical health of the elderly, our solution aims to foster a sense of connection and companionship. Unlike robotics, smart applications, and even VR/AR experiences that offer interaction from a functional or entertainment perspective, AI avatars engage users on a personal level, simulating empathetic conversations and adapting to the emotional states and preferences of its users.

III. MATERIALS, METHODS AND TOOLS

The advanced AI technologies currently being used in elderly care require a comprehensive methodological approach for their analysis, selection, and implementation. These processes rely on proven methods and tools of computer science and artificial intelligence, which also form the basis for avatars that can provide support in care. Data analysis and pattern recognition are at the forefront of the initial analysis phase. By utilizing big data analytics, comprehensive patterns and trends in the health status and behavior of seniors can be identified. These insights are crucial for training AI models that are tailored to the needs of the seniors.

The selection processes focus on AI frameworks that offer adaptive learning capabilities to enable personalized experiences. GPT-based models for natural language processing are preferred to facilitate natural and fluent communication. Moreover, decision trees and other predictive algorithms are significant for providing decision support in care.

The implementation uses machine learning libraries and development environments optimized for working with natural language processing (NLP) and predictive models. Tools like TensorFlow and Keras enable the training of deep learning models, while platforms like OpenAI GPT provide the foundation for developing language AI. Cloud-based services are used for scaling and secure data access, with a strong focus on data protection and compliance with European standards. The use of these tools and methods leads to the development of AI systems that can simulate personal interaction, enable individual learning, and support the cognitive stimulation and well-being of seniors [5].

IV. DESCRIPTION OF THE SOLUTION

The concept elaborated below describes an AI-based assistance system, a family of AI avatars on a tablet, specifically designed to support seniors in

their own homes, acting as an empathetic companion, improving their quality of life [6].

A. Natural Language Processing

The foundation of the system is a speech recognition feature that allows seniors to communicate with the AI in a natural way. Utilizing Natural Language Processing, the system can respond not just to pre-programmed commands but also engage in free-flowing dialogues. This promotes social interaction and takes into account the individual health status and preferences of the users.

B. Learning Abilities and Stimulation

The system adapts and learns from interactions over time. It integrates the user's life story by incorporating biographical information, personal preferences, and family and friendship relationships into the communication and interaction. The assistance system offers a variety of games and cognitive activities specifically designed to enhance the mental abilities of seniors. These range from memory training to problem-solving tasks and are regularly updated to provide stimulation.

C. Network Capabilities

Another integral part is its networking capability. Family members and caregivers can stay in contact with the seniors through the system. This enables quick and easy exchange of information and timely organization of support. In critical situations, the system provides immediate assistance. It is capable of contacting emergency services and also gives regular reminders for daily tasks like taking medication. The architecture is future-oriented and flexible. It can be easily expanded and adapted to a growing number of users to meet the ever-increasing demands.

V. FURTHER ASPECTS AND IMPLEMENTATION

The implementation occurs in several phases, each aiming to improve the living conditions of seniors. A careful analysis of their needs is at the forefront of the project. Experts from the nursing field, psychologists, and technologists work together to ensure seamless integration into the home environment.

A. User-Friendliness

A central focus is on user-friendliness. The AI is designed to be intuitive and operable without prior knowledge. This ensures broad acceptance among seniors, who are expected to interact with the technology. Additionally, the hardware is designed to be robust and low-maintenance, simplifying care by relatives or nursing staff.

B. Data Protection

Furthermore, the system is equipped with a comprehensive data protection strategy. The sensitive data of users are processed and stored with the highest security standards to ensure their privacy. Cognitive stimulation is ensured through regular content updates and the introduction of new activities. Current scientific findings and user feedback are incorporated to continuously improve and adapt the offerings to the needs.

C. Linking Users, Family and Caregivers

For network connectivity, interfaces to common communication platforms are integrated, facilitating easy exchange between seniors and their family members. Caregivers can also access health data through these channels, simplifying monitoring and care.

D. Emergency Response and Behavior Analysis

The emergency response of the system is based on intelligent recognition of deviations in the seniors' normal daily routines. It learns typical behavioral patterns, enabling it to recognize and respond to unusual events that may indicate an emergency. Scalability is ensured through the use of cloud technologies. New functions and services can be centrally implemented and made available to all users without the need for manual updates on-site. The introduction is gradual. Initially, pilot projects with a small user group are conducted to test and further develop the system in practice. User feedback plays a crucial role in refining the system and adapting it to real needs. For long-term support and development, an interdisciplinary team is envisaged, ensuring regular updates and keeping the system in line with the latest technological advancements.

VI. RESULTS AND EXPERIENCES

The focus lies on integrated and forward-looking care. The implementation of technologies such as artificial intelligence, machine learning, and natural language processing enables a personalized user experience. The goal we pursue with the use of the system is not only to improve the quality of life of seniors but also to relieve the burden on nursing staff through more efficient resource utilization. In collaboration with health insurance companies and care facilities, we see the opportunity to realize the following added values.

A. Collaboration and Feedback

Collaboration with health insurance companies will be crucial to understand the framework for the introduction and financing. This will help to develop the system in line with current health regulations and ensure that it is accessible to the broader population. By realizing pilot projects with care facilities, the solution can be tested and

improved in a controlled environment. Such studies are essential to demonstrate the effectiveness of the system and to make specific adjustments to the needs of the users. Practical experience from care facilities will help to further optimize the functionality and user-friendliness.

B. Scalability

The architecture is designed to be easily scalable and adaptable to different care environments. This is particularly advantageous in collaboration with care facilities, as different requirements and contexts need to be considered. In collaboration with health insurance companies and care facilities, data protection and compliance will play a central role from the outset. This is crucial to gain user trust and meet legal requirements. Collaboration with established actors in the healthcare sector will enable the system to access the market more quickly and create a network for future developments and innovations.

C. Long Term Impact Studies

With health insurance companies, long-term impact studies can be initiated to document the benefits of the solution for the health and well-being of seniors over extended periods. The system is still in the development phase. Our expectations are high, but they are based on solid foundations: a deep understanding of the technology, a clear view of the needs of the elderly, and close cooperation with healthcare stakeholders. We expect high user acceptance thanks to its intuitive operation, economic efficiency and personalized interaction. We are aware of the social responsibility that comes with the development of such a system. It is not just about creating a technological product but also about making a contribution that improves the lives of many people and enhances the nursing profession.

VII. CONCLUSION AND FUTURE WORK

The integration of AI-driven approaches in elderly care is still in its infancy, yet expectations are high. Ensuring data security and protecting privacy are central to our considerations, especially due to the sensitive nature of health data. We anticipate the need for ongoing ethical reflection and assessment of the technology to optimize its benefits and address potential risks. Our goal is for AI to be seen not merely as a tool but as a trustworthy partner in health care, constantly interacting with professionals, patients, and relatives. We expect that the future development of AI in healthcare will lead to increasingly individualized and adaptive systems that actively participate, not just react. These systems should have the ability to learn from each interaction and make precise predictions and recommendations through the generalization of learning processes. With the further integration of AI systems into clinical workflows, we foresee a more intense and fruitful collaboration between computer scientists, medical professionals, and caregivers. We are at the threshold of a new era in healthcare, where AI-based assistance systems have the potential to optimize workflows, relieve nursing staff, and improve senior care. In conclusion, we emphasize that despite the excitement for technological advancements, the human aspect remains essential. Technology should complement and support human work, but the core aspects of care – empathy, understanding, and personal attention – must not be neglected.

REFERENCES

[1] Topol, E. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.

[2] Ramesh, A. N., Kambhampati, C., Monson, J. R. T., & Drew, P. J. (2004). Artificial intelligence in medicine. Annals of the Royal College of Surgeons of England, 86(5), 334-338.

[3] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology, svn-2017.

[4] Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. Journal of the American Medical Association, 319(13), 1317-1318.

[5] Meskó, B., Hetényi, G., & Győrffy, Z. (2018). Will artificial intelligence solve the human resource crisis in healthcare? BMC Health Services Research, 18(1), 545.

[6] Luxton, D. D. (2014). Artificial intelligence in psychological practice: Current and future applications and implications. Professional Psychology: Research and Practice, 45(5), 332.

CAD Tool for Breast Cancer Prediction Using Multiple Deep-learning Models

Maura Mengoni
Department of Industrial Engineering and Mathematical Science
Polytechnic University of Marche,
Ancona, Italy
e-mail: m.mengoni@univpm.it

Aubudukaiyoumu Talipu, Giampiero Cimini, Marco Luciani
Technical Department
EMOJ S.r.l.
Ancona, Italy
e-mail: a.talipu@emojlab.com, g.cimini@emojlab.com, m.luciani@emojlab.com

Luca Giraldi
Department of Economy
University of Macerata,
Macerata, Italy
e-mail: luca.giraldi@unimc.it

Mauro Savino
Research and Development
Diatech Pharmacogenetics S.r.l.
Jesi, Italy
e-mail: mauro.savino@diatechpharmacogenetics.com

***Abstract*—Breast cancer is one of the leading causes of cancer death among women worldwide. It represents a global health concern due to the lack of effective therapeutic regimens that could be applied to all breast cancer patients. Breast cancer treatment decisions rely on clinicopathologic parameters. However, this approach is replete with limitations as it fails to define prognosis uniquely and is not always sufficient to settle unequivocally on the best type of treatment for breast cancer patients. The molecular diagnostic efforts have been focused mainly on Estrogen Receptor (ER)-positive (Luminal A) breast cancer being the most represented breast cancer subtype (70% of patients) with a standard treatment (endocrine therapy for five years) and a good prognosis. However, at least 20% of patients will suffer a distant recurrence within ten years. Although many molecular tests have been developed to identify the patients at risk of recurrence, a definite, reliable and effective in vitro diagnostic device that stratifies patients at high risk and low risk of relapse, directing therapeutic decisions, is still a significant clinical need. This study aims to fill this gap by investigating and developing a new approach for better stratification of breast cancer patients in the risk categories of recurrence. It is based on the integration of clinical and digital pathology analysis. The combined analysis, indeed, aims to further categorize the patients with an intermediate risk of recurrence either in the low-risk group with no necessity of chemotherapy or in the high-risk group that needs chemotherapy. The paper presents the approach, the implemented Computer-Aided Diagnosis (CAD) tool and finally, the results of evaluating its predictive accuracy. The tool achieved 88% accuracy in histological image classification, 95% in cancer grade prediction and 71% in 10-year recurrence prediction.**

Keywords: breast cancer; Computer-Aided Diagnosis; Histopathological Imaging; Artificial Intelligence

I. INTRODUCTION

Breast cancer is the most common type of cancer worldwide and the leading cause of death among women [1]. It is worth noticing that early detection and timely diagnosis of breast cancer are of vital importance in saving lives.

Cancer screening helps detect cancer or precancerous abnormalities in individuals with no symptoms. The primary goal of cancer screening is to identify cancer at an early stage when it is more treatable, potentially leading to better outcomes and increased chances of survival. Currently, histopathological tissue analysis by a pathologist represents the only definitive method for confirmation of the presence or absence of disease and disease grading or the measurement of disease progression [2].

Histopathology slides provide a comprehensive view of disease and its effect on tissues because the preparation process preserves the underlying tissue architecture. As such, some disease characteristics may be deduced only from a histological image. However, the histological image analysis process is tedious and subjective, causing inter-observer variations even among senior pathologists [3]. With the advancement of computer vision and image processing based on deep learning algorithms, Computer-Aided Diagnostic (CAD) systems can overcome these difficulties. It can extract the essential information from the histological images and detect patterns not visible to the human eye [4].

Another crucial field of research area related to breast cancer is the prediction of breast cancer recurrence. About 80% of patients initially presenting with early-stage disease have a recurrence in 5 years, and 30% of patients have a recurrence of cancer within 10 years after the completion of initial treatment [1]. The risk of recurrence is a significant concern for individuals who have undergone treatment for cancer. Various factors, i.e., the stage of initial cancer, specific biological markers (such as hormone receptor status), the

effectiveness of the initial treatment, and individual patient characteristics, are considered to influence the risk of recurrence [1]. With clinicopathologic characteristics of cancer patients, it is possible to predict 5-year cancer recurrence [5]. Doctors could use such a prediction to make a tailored treatment plan.

CAD systems leverage deep learning and multidisciplinary knowledge and techniques to analyze medical imaging and non-imaging data and provide the analyzed results to clinicians as second opinion or decision support in the various stages of the patient care process [6].

CAD tools, such as Aiforia, PathAI, Adjuvant!, PREDICT, and CanAssist-Breast (CAB) are among the popular ones. Aiforia [7] and PathAi [8] have similar capabilities, including automated image analysis, quantification of pathology features, and pattern recognition. Adjuvant! [9] and PREDICT [10] online tools are widely used in breast cancer recurrence prediction. Adjuvant! does not produce accurate results and is no longer available online [11]. PREDICT utilizes patient and tumor characteristics to generate predictions for individual patients. It helps clinicians and patients make informed decisions about treatment by estimating the likelihood of recurrence. Population-based study [12] conducted on older patients reported the effectiveness of PREDICT in 5-year recurrence and a slight overestimation in 10-year recurrence prediction. CAB [13] is another promising tool for the immunohistochemistry-based prognostic test; it utilizes biomarkers and clinical parameters, such as tumor size, grade and node status as inputs to generate a risk score and categorizes patients as low- or high-risk for distant recurrence within 5-years of diagnosis.

All the tools studied can perform only one histological image analysis or recurrence prognostics, not both. However, they have excellent 5-year recurrence results, while 10-year recurrence remains challenging.

The primary purpose of this research is to create a diagnostic CAD tool that can detect cancerous and non-cancerous areas in a breast cancer histological image, to predict cancer staging and to develop a generalized estimation of the risk of breast cancer 10-year recurrence, by combining the histological and clinical patient data. The paper describes in detail the methodology adopted for its development and the validation results. It is organized as follows: Section 2 shows background information and related work in machine learning and CAD tools; Section 3 describes the methodology adopted; Section 4 presents the experiments conducted and best results achieved; Section 5 discusses the experiment results and findings; and Section 6 concludes the paper by summarizing and providing future directions of work.

II. RELATED WORK

Machine learning has significantly advanced CAD tools in various ways, particularly in medical imaging, including breast cancer detection and diagnosis. CAD tools developed with conventional machine learning methods mainly use hand-engineered features based on the domain knowledge and expertise of human developers who translate the perceived image characteristics to descriptors that mathematical functions or conventional image processing techniques can implement. The recent advancement in computing power and dataset sizes allowed the application of deep convolutional neural networks (DCNN) to image classification problems. Contrary to the traditional approach of hand-crafted feature extraction methods, DCNNs learn useful features directly from the training image patches by optimizing the classification loss function.

Several studies focused on histological images with DCNNs. The works range from pioneering studies that introduced the concept of using deep learning for breast cancer diagnosis to sophisticated architectures tailored for specific tasks like segmentation [14] and feature extraction [15].

DCNN-based CAD systems can automatically extract meaningful features from histological images. These features include texture, color, shape, and intensity patterns. They can also perform image segmentation, which involves identifying and delineating specific regions of interest within the histological images. It is beneficial for isolating cancerous lesions or specific cell types. Usually, the cancer diagnosing process using a histological image consists of the following steps [16]. Firstly, tissue specimens are extracted through biopsy, affixed on glass slides, and stained with hematoxylin and eosin (H&E). Then, an expert histopathologist examines the glass slides under a light microscope to provide the diagnosis for each sample. Accurate interpretation of glass slides is crucial to avoid misdiagnoses, which require extensive time and effort by the pathologist. Each person could have up to a dozen biopsy samples that require analysis. It displays the necessity of computational digital pathology to augment and automate diagnosis processes by scanning digitized whole slide images (WSI). WSI contains many cells; the image could consist of tens of billions of pixels, which is usually hard to analyze. However, resizing the entire image to a smaller size, such as 256 X 256, would lead to the loss of information at the cellular level, resulting in a marked decrease in identification accuracy. Therefore, the entire WSI is commonly divided into partial regions of about 256 X 256 pixels (“patches”), and each patch is analyzed independently.

Araujo et al. [17] proposed a deep convolutional neural network combined with an SVM (support vector machine) to classify hematoxylin and eosin (H&E) stained histological images and achieved accuracies of 77.8% for four class (normal tissue, benign lesion, in situ carcinoma and invasive carcinoma) classification and 83.3% for two class (carcinoma, non-carcinoma) classification.

In [18], the popular DCNNs architectures pre-trained on ImageNet, such as VGG, ResNet and Inception extract the essential features from the images, and then gradient-boosted trees classifier is applied to classify the images.

Ensemble approaches like [15] and [19] also took similar approaches, except they employed an ensemble of DCNNs, namely VGG19, MobileNetV2 and DenseNet201, to extract visual features and then applied the boosting framework to achieve superior results in the detection of cancerous and non-cancerous areas from the histological image.

CAD tools, such as Aiforia and PathAI have cancerous and non-cancerous area detection features from a given histological image. Although breast cancer is detected early and the treatment is started soon after diagnosis, the cancer cells remain in the body undetected; after a certain period, it may recur. Machine learning methods are also applied to advance the prediction accuracy in breast cancer recurrence prediction. Usually, the datasets contain many features, which may mislead the prediction process as some features may lead to confusion or inaccurate prediction [20].

Feature selection is an essential first step in breast cancer recurrence prediction. In [20], a hybrid multi-stage learning technique based on brain-storming optimization was applied to study the most effective features, and it concluded that the feature selection is highly dependent on the applied classification algorithm and dataset used. [21] studied clinicopathologic characteristics of 579 breast cancer patients. It used statistical feature selection and particle swarm optimization to select and refine important features. It compared SVM, Decision Tree (DT) and Neural Network classifiers to predict breast cancer 5-year recurrence. It used the local invasion of the tumor, the number of tumors, the number of metastatic lymph nodes, the histological grade, the tumor size, estrogen receptor, and lympho-vascular invasion. PREDICT online tool predicts the recurrence based on features, such as breast cancer type, patient age, menopause, ER status, Ki-67, tumor size and tumor grade.

Current existing CAD tools perform only one of the different stages of the diagnostic process; obtaining a general all-in-one diagnostic report requires the involvement of different tools, which is not an efficient workflow. The main contribution of this work is to develop an all-in-one CAD tool that utilizes machine learning algorithms like DCNNs and eXtreme Gradient Boosting (XGBoost). The developed CAD tool can generate a full breast cancer diagnostic and prognostic report by combining histological image analysis and clinical and histological data analysis.

III. METHODOLOGY

The development of the novel CAD tool consists of the following steps: data collection, dataset creation (stain normalization, patch extraction), model training and validation steps.

A. Data collection

Histological image analysis faces data variability, class imbalance, and potential bias challenges in general. Ensuring a representative and diverse dataset is crucial for training supervised DNN models that generalize well to real-world scenarios. It directly affects the performance of the trained model on new, unseen images. At the same time, accurate labels that indicate the presence or absence of the target condition, such as cancerous or non-cancerous tissue, are essential.

The data collection is achieved by digitizing the samples collected from anonymous patient biopsy slides provided by Verona Borgo Trento Hospital (Italy) with NED DP digital microscope. 300 sets of histological images with various magnification levels, specifically, 1.25x, 2x, 4x, 10x, 20x, and 40x, are collected and manually labelled by the Verona Borgo Trento hospital medical practitioners.

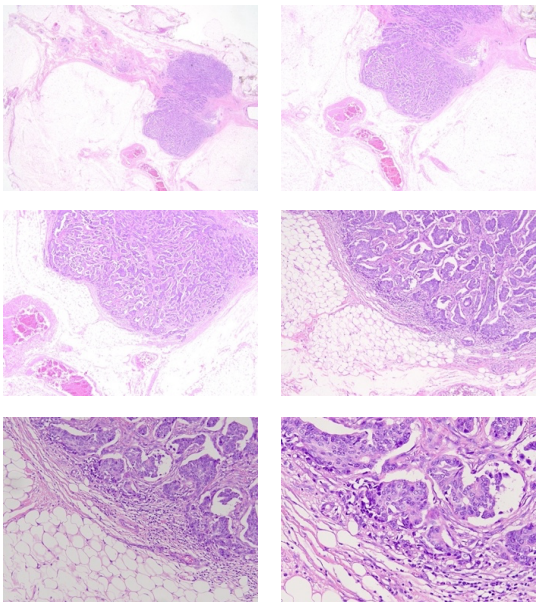


Figure 1. Different magnification variations (Starting from top left 1.25x, 2x, 4x, 10x, 20x, 40x)

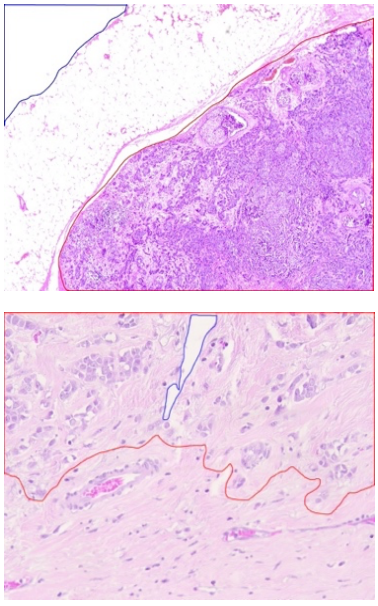


Figure 2. Labelled histological images

Each image has 1640x1175 resolution, and the labels are defined as blue and red color-coded lines on top of the tissue image, indicating areas without tissue cells and with cancerous tissue cells, respectively. As indicated in the following sample image, a closed-shape label makes it easy to separate and identify the areas in the next phase. Every image is also accompanied by clinical information and features, they are listed in the following table.

TABLE I. ALL FEATURES FROM CLINICAL DATA

All Features		
pT	HER2	RECIDIVA
Numero LN metastici	PR	tipo di recidiva
pN	N NPI	tempo di recidiva (mesi)
Grado	NPI SCORE	Follow up mesi
STADIO	NPI GROUP	Luminal
DIAMETRO MM	Adiuvante	Età alla diagnosi
ISTOTIPO	Overall survival (mesi)	Menopausa
Ki67	DOA	ER

B. *Dataset creation and model training*

The histological images are often stained to enhance the visibility of structures and cells. Variations in staining procedures can lead to differences in color and intensity, challenging comparing images or applying consistent analysis.

Firstly, considering the visual consistency and reproducibility of the experiments, stain normalization technique proposed in [22] is applied to the histological images to normalize the stains. Then, multiple datasets are created based on magnification levels with different image sizes. The clinical data was recorded by different medical specialists, each using a different structure and labelling. A preprocessing method standardizes the labels and filters out the missing data samples. The histological images without complete clinical data are discarded to prevent mismatching in the result. Secondly, the specific algorithms applied are chosen. The classification of cancerous and non-cancerous areas from the image is considered a binary classification, and the patch based DCNN approach is the best suited. The literature study shows that DCNN architectures like VGG and Inception pre-trained on ImageNet resulted in highly accurate classification models; therefore, fine-tuning the pre-trained models is favored.

A combination of multiple input sizes (patch dimension), various learning rates and batch sizes are experimented with, and the best accuracy model is selected at the end. Table II shows the specific parameters experimented during the model training. XGBoost is a popular machine learning algorithm known for its efficiency, speed, and performance in various predictive modelling tasks. After selecting the features with statistical feature selection, XGBoost is used in grade prediction.

TABLE II. VARIABLES EXPERIMENTED IN TRAINING

Architectures	VGG16	VGG19	Inception
Patch dimensions	64x64	150x150	200x200
Batch size	32	64	128
Learning rate	0.01	0.001	0.003

Finally, breast cancer recurrence is predicted with linear regression with the selected features. The experiment and result section describes the testing results regarding the parameters and model training.

IV. EXPERIMENTS

The experiment consists of preprocessing, image classification, grade prediction and recurrence prediction steps. In the experiments conducted, the images are pre-processed, multiple datasets are created by extracting multi-dimensional patches from them. Then, histological image classification, grade prediction and recurrence prediction models are trained. Combinations of various parameter are experimented, model accuracies and algorithms used are reported. The models are integrated to the CAD tool developed.

A. *Data preprocessing*

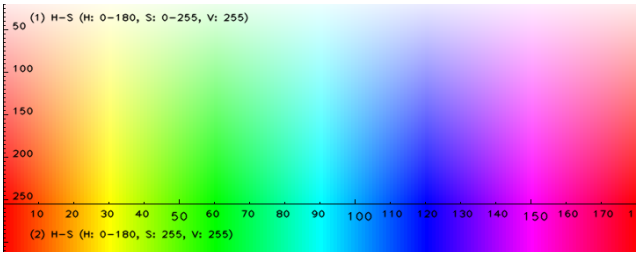


Figure 3. HSV color spectrum

In preprocessing, the color-coded image labels are separated with computer image processing techniques. With HSV color map spectrum (in Figure 3) and OpenCV, the blue and red color masks are created to separate the corresponding labels in the histological image. The preprocessing and construction of datasets are illustrated in Figure 4.

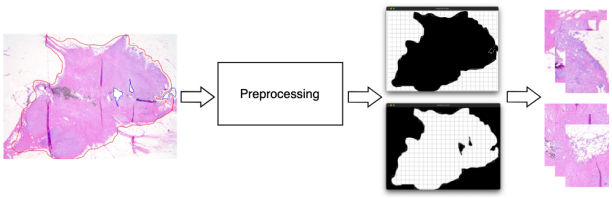


Figure 4. Preprocessing

Dilatation and erosion [23] techniques are also implemented to enhance the label continuity and fully surround the area of interest. Areas (in Figure 5) labelled with blue labels are excluded to minimize the impact of false classification. After successfully separating the color-coded labels, areas with negative and positive labels are represented with black and white masks. Then, patches are extracted from the corresponding areas to construct datasets.

A combination of 64x64, 150x150 and 200x200 patch sizes and magnification of 1.25x, 2x, 4x, 10x, 20x, and 40x are used to construct multiple datasets. The patch from the

edges contains positive and negative areas. A threshold of 0.7 is applied to label them. This threshold value is considered better suited because it produces relatively balanced datasets to work with. For instance, if 70% of the region is from a positively labelled region, it is labelled with a positive label.

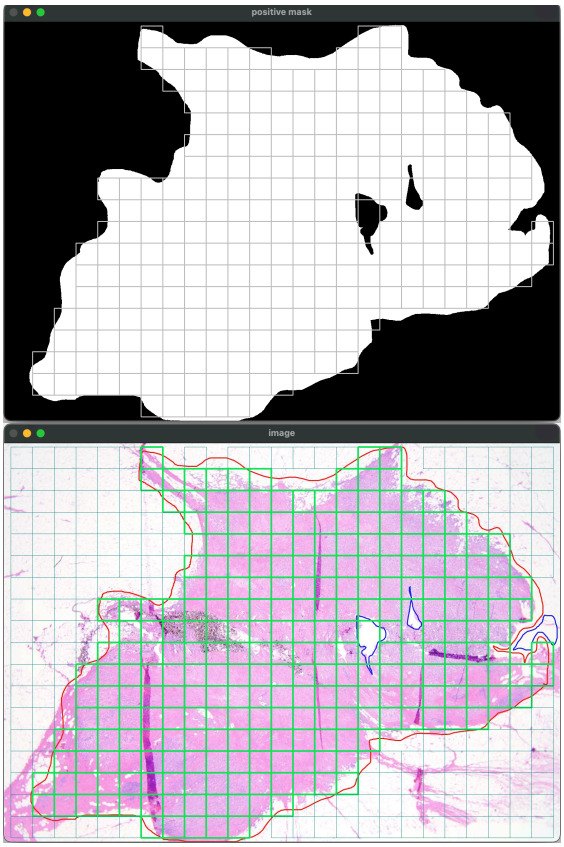


Figure 5. Label masks and patch extration

AWS Glue DataBrew [24] service is applied to establish a homogeneous dataset with corresponding clinical data and to facilitate the creation of an automated data pipeline, streamlining the data ingestion and preprocessing. Leveraging AWS SageMaker [25], essential data cleaning and preprocessing steps, such as feature scaling, categorical data encoding, and augmenting are conducted. To ensure the consistency of all the clinical data rows, “ISTOTIPO” is split into a set of derived variables namely, “ISTOTIPO_CDI”, “ISTOTIPO_CLI”, “ISTOTIPO_NST”, “ISTOTIPO_TUBULARE”, “ISTOTIPO_LOBULARE”, “ISTOTIPO_APOCRINO”, “ISTOTIPO_MICROPAPILLARE”, “ISTOTIPO_MUCINOSO”, “ISTOTIPO_PAPINCAPS” using one-hot vector encoding. This method provided a more comprehensive representation of the original “ISTOTIPO” feature and enriched the dataset, enhancing the model capacity for generating more accurate and insightful predictions. As a result of the preprocessing, a dataset of 300 data samples, each with 26 attributes, is constructed.

B. *Histological Image Classification*

The datasets constructed in the previous step are used to train DCNN models. Training, validation, and testing splits of 6:3:1 and 7:2:1 are experimented. Among all the trials conducted with different combinations of batch size, learning rate, input size (patch size) and DCNN architecture, fine-tuning pre-trained VGG16 on ImageNet with the following parameters (Table III) resulted in the best accuracy model.

TABLE III. THE BEST ACCURACY MODEL PARAMETERS

Architecture	VGG16
Magnification	40
Patch dimension	200x200
Batch size	128
Learning rate	0.001
Accuracy	87.6%
F1	0.88

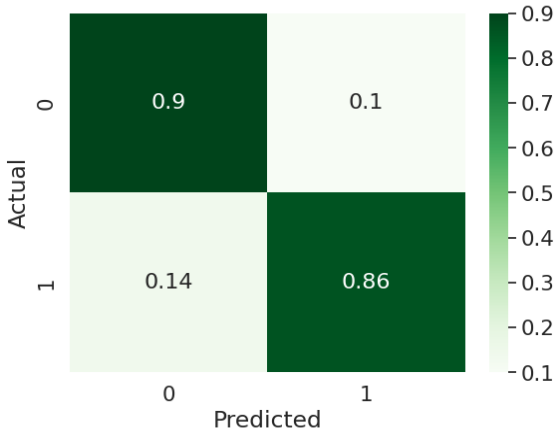


Figure 6. Label masks and patch extration

The confusion matrix of the model validation prediction is shown in Figure 6.

C. *Grade prediction*

The training process for this task revolved around a multiclass classification problem, where the goal was to categorize data into one of multiple predefined classes or categories. In multiclass classification, multiclass categorical cross entropy metric quantifies the differences between the predicted class probabilities and the true class labels for each data point. The features selected to train the model are listed in the Table IV below.

The model trained with XGBoost, achieved a validation accuracy of 95%. The multiclass logarithmic loss for the corresponding model is 0.21. The confusion matrix (in Figure 7) provides an in-depth insight into the model prediction, revealing its effectiveness in multiclass classification.

TABLE IV. ALL FEATURES FROM CLINICAL DATA

Grado	ISTOTIPO_APOCRINO	ADIUVANTE_CHT
pT_adjusted	ISTOTIPO_LOBULARE	ADIUVANTE_RT
Numero LN metastatici_adjusted	ISTOTIPO_TUBULARE	ADIUVANTE_CT
pN_adjusted	ISTOTIPO_NST	ADIUVANTE_OT
STADIO_adjusted	ISTOTIPO_CLI	Follow up mesi
DIAMETRO MM	ISTOTIPO_CDI	LUMINAL_adjusted
ISTOTIPO_PAPINCAPS	Ki67	età alla diagnosi
ISTOTIPO_CRIBRI	PR	menopausa
ISTOTIPO_MUCINOSO	N NPI	ER
ISTOTIPO_MICROPAPILLARE	NPI SCORE	% cellule neoplastiche

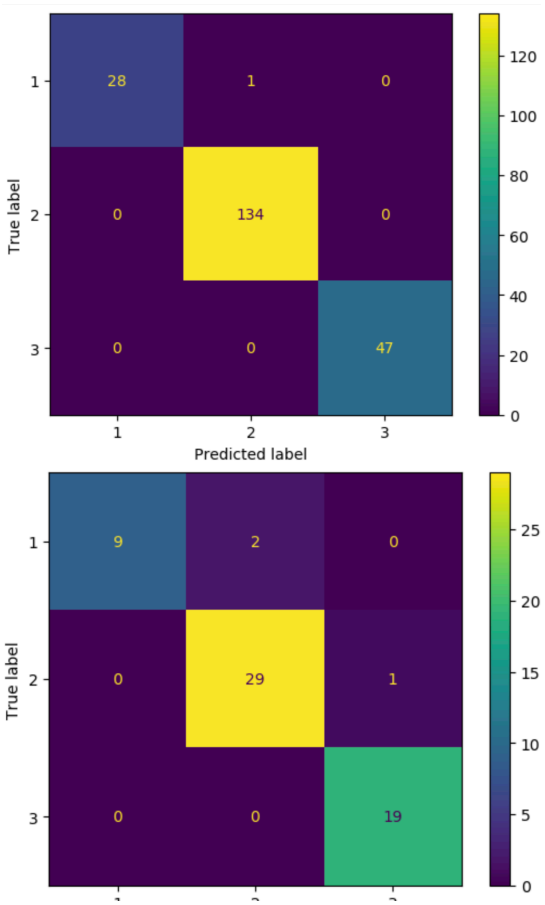


Figure 7. Label masks and patch extration

D. Breast cancer recurrence prediction

The correlations between features are explored and reduced to obtain a maximum accuracy model. However, because of the ambiguity in the dataset recurrence months, many rows are discarded, and the multiple regression model is trained with very few data, around 40 records. The final accuracy obtained with the multiple linear regression for 10 - year recurrence prediction is 71%.

V. DISCUSSION

The accuracies achieved by image classification and linear regression are lower than to state-of-the-art results, especially with the linear regression. Fine tuning the pretrained VGG16 model achieved 87.6% accuracy, a reasonably good result but further investigations need to be conducted to improve the model robustness and accuracy. A comparison study should be conducted using open datasets to compare and validate the achieved classification result. The grade prediction with XGBoost algorithm achieved 95% accuracy, it demonstrates the effectiveness and efficiency of XGBoost algorithm. Finally, the linear regression for predicting the breast cancer 10-year recurrence only achieved 71% accuracy. Applying different machine learning algorithms, such as DT or SVM could improve the accuracy further. Regarding the data, greater magnification levels, such as 100x, 200x with more data certainly improve the over-all accuracies obtained in this study. Different subsets of features could be investigated in XGBoost and multiple linear regression to further improve the model accuracies. Moreover, different approaches, such as fusing multiple DCNNs as a feature extractor and combining different types of classifiers, such as SVM, DT or XGBoost could be the path to achieve a better result.

VI. CONCLUSION AND FUTURE WORK

The work presented in this paper aims to create an all-in-one breast cancer diagnostic tool for (ER)-positive breast cancer patients. The histological image classification by fine-tuning ImageNet pretrained VGG16 model obtained 88% accuracy, the cancer grade prediction with XGBoost algorithm achieved 95% accuracy, and the cancer recurrence prediction with linear regression resulted 71% accuracy. It is an essential initial step in our future study direction. Histological image analysis and clinical data analysis are combined in the proposed CAD tool to predict breast cancer recurrence. This type of CAD tool is very useful in assisting doctors to reduce their workload and improve the reproducibility of breast cancer diagnostics.

Future studies will improve the accuracies and robustness of the models, acquire further labelled data and test with different DL approaches. Fusing molecular and genetic data and imaging feature might also enable a comprehensive understanding of disease characteristics.

REFERENCES

[1] C. Mazo, C. Aura, A. Rahman, W. M. Gallagher, and C. Mooney, “Application of Artificial Intelligence Techniques to Predict Risk of Recurrence of Breast Cancer: A Systematic Review,” *J Pers Med*, vol. 12, no. 9, pp. 1-11, 2022, doi: 10.3390/jpm12091496.

[2] M. N. Gurcan et al. “Histopathological Image Analysis: A Review,” *IEEE Rev Biomed Eng*, vol. 2, pp. 147–171, 2009, doi: 10.1109/RBME.2009.2034865.

[3] J. G. Elmore *et al.*, "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *JAMA - Journal of the American Medical Association*, vol. 313, pp. 10-11, 2015, doi: 10.1001/jama.2015.1405.

[4] S. Robertson, H. Azizpour, K. Smith, and J. Hartman, "Digital image analysis in breast pathology—from image processing techniques to artificial intelligence," *Translational Research*, vol. 194. pp. 20 2018. doi: 10.1016/j.trsl.2017.10.010.

[5] A. M. Gonzalez-Angulo *et al.*, "High risk of recurrence for patients with breast cancer who have human epidermal growth factor receptor 2-positive, node-negative tumors 1 cm or smaller," *Journal of Clinical Oncology*, vol. 27, pp. 33-34, 2009, doi: 10.1200/JCO.2009.23.2025.

[6] H. P. Chan, R. K. Samala, and L. M. Hadjiiski, "CAD and AI for breast cancer - Recent development and challenges," *British Journal of Radiology*, vol. 93, no. 1108. pp. 1-2, 2020. doi: 10.1259/bjr.20190580.

[7] "Aiforia." Accessed: 01.2024. [Online]. Available: https://www.aiforia.com/

[8] "PathAI." Accessed: 01.2024. [Online]. Available: https://www.pathai.com/

[9] P. M. Ravdin *et al.*, "Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer," *Journal of Clinical Oncology*, vol. 19, no. 4, 2001, doi: 10.1200/JCO.2001.19.4.980.

[10] "PREDICT." Accessed: 01.2024. [Online]. Available: https://breast.predict.nhs.uk/

[11] N. A. De Glas *et al.*, "Validity of adjuvant! Online program in older patients with breast cancer: A population-based study" *Lancet Oncol*, vol. 15, no. 7, pp. 1, 2014, doi: 10.1016/S1470-2045(14)70200-1.

[12] N. A. De Glas *et al.*, "Validity of the online PREDICT tool in older patients with breast cancer: A population-based study," *Br J Cancer*, vol. 114, no. 4, pp. 1-2, 2016, doi: 10.1038/bjc.2015.466.

[13] M. M. Bakre *et al.*, "Clinical validation of an immunohistochemistry-based CanAssist-Breast test for distant recurrence prediction in hormone receptor-positive breast cancer patients," *Cancer Med*, vol. 8, no. 4, pp. 1755–1764, Apr. 2019, doi: 10.1002/cam4.2049.

[14] L. Yang, P. Meer, and D. J. Foran, "Unsupervised segmentation based on robust estimation and color active contour models," *IEEE Transactions on Information Technology in Biomedicine*, vol. 9, no. 3, 2005, doi: 10.1109/TITB.2005.847515.

[15] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Classification of Histopathological Biopsy Images Using Ensemble of Deep Learning Networks," *CASCON 2019 Proceedings - Conference of the Centre for Advanced Studies on Collaborative Research - Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*, pp. 92–99, Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.11870

[16] M. Veta, J. P. W. Pluim, P. J. Van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: A review," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5. IEEE Computer Society, pp. 1400–1411, 2014. doi: 10.1109/TBME.2014.2303852.

[17] T. Araujo *et al.*, "Classification of breast cancer histology images using convolutional neural networks," *PLoS One*, vol. 12, no. 6, Jun. 2017, doi: 10.1371/journal.pone.0177544.

[18] D. M. Vo, N. Q. Nguyen, and S. W. Lee, "Classification of breast cancer histology images using incremental boosting convolution networks," *Inf Sci (N Y)*, vol. 482, 2019, doi: 10.1016/j.ins.2018.12.089.

[19] K. Das, S. P. K. Karri, A. Guha Roy, J. Chatterjee, and D. Sheet, "Classifying histopathology whole-slides using fusion of decisions from deep convolutional network on a collection of random multi-views at multi-magnification," in *Proceedings - International Symposium on Biomedical Imaging*, 2017. doi: 10.1109/ISBI.2017.7950690.

[20] M. Alwohaibi, M. Alzaqebah, N. M. Alotaibi, A. M. Alzahrani, and M. Zouch, "A hybrid multi-stage learning technique based on brain storming optimization algorithm for breast cancer recurrence prediction," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, 2022, doi: 10.1016/j.jksuci.2021.05.004.

[21] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas, and F. Mokarian, "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning," *Comput Struct Biotechnol J*, vol. 15, 2017, doi: 10.1016/j.csbj.2016.11.004.

[22] M. Macenko *et al.*, "A method for normalizing histology slides for quantitative analysis," in *Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009*, 2009. doi: 10.1109/ISBI.2009.5193250.

[23] "Eroding and Dilating." Accessed: 01.2024. [Online]. Available: https://docs.opencv.org/3.4/db/df6/tutorial_erosion_dilatation.html

[24] "AWS Glue DataBrew", Accessed: 01.2024. [Online]. Available: https://aws.amazon.com/glue/features/databrew/

[25] "AWS SageMaker", Accessed: 01.2024. [Online]. Available: https://aws.amazon.com/sagemaker/

Human-AI Collaboration Cycle in the Development Stage of An AI-enabled System

Pi-Yang Weng
Dept. of Management Info Systems
National Chengchi University
Taipei, Taiwan
email: piyangong@gmail.com

Rua-Huan Tsaih
Dept. of Management Info Systems
National Chengchi University
Taipei, Taiwan
email: tsaih@mis.nccu.edu.tw

Hsin-Lu Chang
Dept. of Management of Info Systems
National Chengchi University
Taipei, Taiwan
email: hlchang@mis.nccu.edu.tw

***Abstract*—Explainable Artificial Intelligence (XAI) has garnered attention in the AI system development in recent years, especially in the high-stakes decision scenarios, such as medical and healthcare domains. In this paper, we present a framework named Human-AI Collaboration Cycle. The framework emphasizes the collaboration between domain experts and AI system in the development stage of an AI-enabled system through an introduction of XAI. We propose that the introduction of XAI can enhance domain experts' engagement in the stages of model evaluation and validation, then further review and engage in the data preprocessing, which in turn, improves their comprehensibility and trust toward the system. To validate our framework, we will conduct a field experiment in a hospital, in which nurses, as domain experts, and AI engineers will work together to develop an AI-enabled fall detection system with model explainability. We will evaluate the role of Local Interpretable Model-agnostic Explanations (LIME), one of the noted XAI tools, in the proposed Human-AI Collaboration Cycle.**

Keywords-engagement; domain expert; XAI; comprehensibility; trust.

I. Introduction

Organization for Economic Cooperation and Development (OECD) has promoted the concept of the responsible AI, suggesting that AI-related actors need to advocate human-center value, transparency, explainability, and accountability [1]. In order to obtain a better output performance, past research suggests that domain experts are required to engage in the Machine Learning (ML) pipeline to assist in building an AI-enabled system [2]. It is also important to have domain experts kept in the loop to optimize the ML model [3]. However, a system developed by AI technology is not usually based on a clear statistics and probability theory. It is inevitable for domain experts to consider it as a black box even though its inputs and outputs are useful mappings. Therefore, it is necessary that machine learning and AI systems need to be explainable and comprehensible in human terms, which is instrumental for validating the quality of an AI system outputs [4]. The output of the black box needs a reasonable explanation for domain experts to trust in the AI-enabled system. In response to this issue, XAI has been more widely recognized in recent years.

It is essential that domain experts increase their trust in the AI-enabled system and further optimize the AI model and adopt it during Human-AI Collaboration (HAC). During HAC, a better output performance is expected through the domain experts' engagement in the higher quality training data generation [2]. Therefore, domain experts need to engage in the data preprocessing, such as data cleaning, data labeling, and feature selections. In recent years, AI Model Explainability has been receiving greater attention as well. However, user trust is not easy to build due to lack of transparency, especially in high-risk decision contexts, such as medical and healthcare domains [5]. We will present an XAI tool to unveil the black box to build user trust in an AI-enabled system.

This research findings will provide AI-enabled system designers with a Human-AI Collaboration Cycle framework as a guideline for developing a responsible AI system. Also, this research will highlight the importance of domain experts' engagement in the ML pipeline in the development stage of an AI-enabled system and highlight the functionality of XAI incorporated in the model evaluation/validation process, which could enhance user trust in an AI-enabled system.

In Section 2, we reviewed current concepts on Human-AI collaboration, ML pipeline, and XAI. In Section 3, we proposed a conceptual model named Human-AI Collaboration Cycle. In Section 4, we proposed a research methodology with IT Artifact, Hypotheses, and Experiment Design to validate our framework. In Section 5, we made a preliminary conclusion for this research and proposed our future work.

II. Literature Review

The literature review of this research will be composed of three parts: Human-AI Collaboration (HAC), Machine Learning Pipeline and Explainable AI (XAI).

A. Human-AI Collaboration (HAC)

Human experts and AI have different yet complementary capabilities by which they can work together to have an effective decision-making [6]. AI is not just a tool; it may become a teammate to enhance team performance [7]. Humans and AI can have mutual learning through which AI or algorithms can learn from humans and humans can acquire insights from AI or algorithms [8]. ML needs methods that engage domain experts directly into the ML process and have them in the loop until the desired results are received [2]. After building an AI

model, data scientists often need to find a domain expert to help interpret the test results and validate whether they make sense or not [9].

Humans and AI can work together as a symbiotic system through which humans can gain intelligence augmentation and AI can learn from humans' feedback through interactions [10]. AI system designers could consider a human-AI team building based on the core competencies brought in by humans and the core capabilities of the AI teammates [7]. Therefore, it is required that domain experts need to collaborate with AI systems through AI engineers in the development stage of an AI system in order to obtain a better system performance.

B. Machine Learning Pipeline

The ML pipeline starts with data extraction and analysis and then obtains a trained model after model evaluation and validation [11]. The pipeline is shown in Figure 1.

The key tasks for each ML step are described as follows:

- Data extraction and analysis
  Select and integrate the relevant data from various data sources for the ML task. Also, identify the data preparation and feature engineering that are needed for the model.
- Data preparation
  It involves data cleaning and data splitting into training data and test data.
- Model training
  The data scientist implements different algorithms with the prepared data to train various ML models.The output of this step is a trained model.
- Model evaluation and validation
  The model is validated on a holdout set to evaluate the model quality. The output of this step is a set of metrics to assess the quality of the model.

The domain experts need to join the training data labeling task, in the case of supervised learning, for obtaining high-quality training datasets and avoiding garbage in, garbage out results [12]. Also, they are required to engage in the model evaluation and validation [2]. Before a trained model is accepted by the domain experts, usually the system users, they are required to stay in the ML pipeline, especially in the stages of data extraction and analysis and model evaluation back and forth.

C. Explainable AI (XAI)

XAI is a useful tool to unveil the ML black box and provides an explanation for each AI system output [13]. XAI is especially instrumental in medicine and healthcare to ensure that the AI system outputs produced by the AI model are correct and justifiable [14]. It is necessary to explain the AI system's decision to increase the users' trust in the system. If AI system users can clearly understand the particular reasons for each system output, they will tend to trust in the AI system [15].

As domain experts have more understanding on the AI algorithm and the explanation for each system output, they tend to provide more feedback on the data preprocessing, such as data extraction and feature selections, further engaging in the ML pipeline and, hence, contributing their domain knowledge into the AI model refinement. Therefore, with XAI incorporated, domain experts will stay in the loop of ML pipeline until an acceptable AI model is achieved.

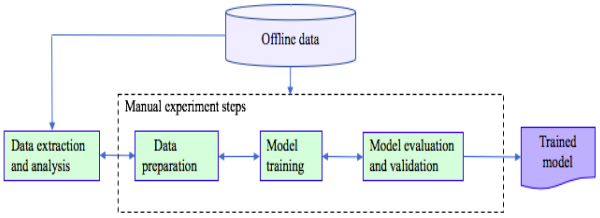


Figure 1. Machine learning pipeline.

III. Conceptual Model

As domain experts are the key AI-enabled system users, the HAC requires domain experts' engagement in the development stage for building a high-quality training dataset and achieving a better model for deployment. Moreover, AI system users will enhance their comprehensibility with the AI model by incorporating XAI into the model evaluation and validation process [16]. With this comprehensibility, AI system users will have greater trust in the AI system and, therefore, adopt the system.

It is possible that the newly trained model would fail in the next few tests before deployment. One of the possible reasons is that the model does not cover some real-world cases, which may be attributed to the introduction of XAI. With XAI, it could facilitate domain experts' engagement in the model evaluation and validation with new test data. Hence, the domain experts will need to re-engage in the ML pipeline for the training dataset review. Therefore, it constructs a cycle in the HAC. We coin it as Human-AI Collaboration Cycle, which is shown in Figure 2.

There are four components in the HAC cycle:

- Data Engagement
  In this research, Data Engagement refers to domain experts' engagement in the data preprocessing including data extraction, data cleaning, data labeling, and feature selections. With domain experts' engagement, the training data would have higher quality to train a better model. Therefore, data engagement in ML pipeline implies that domain experts would have partial responsibility for a better trained model.
- User Comprehensibility
  XAI is a technology tool to unveil the black box, which could help domain experts comprehend the model and algorithm logic. LIME, as one of XAI tools, will present some key features for each instance, i.e., each input [13]. Its output format is shown in Figure 3.

  In order to measure the speed of human movement, we use a human skeleton marked with four key coordinates [17], as shown in Figure 4. Point 1 to Point 4 represents

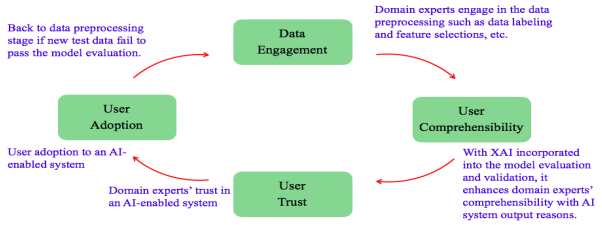


Figure 2. Human-AI collaboration cycle.

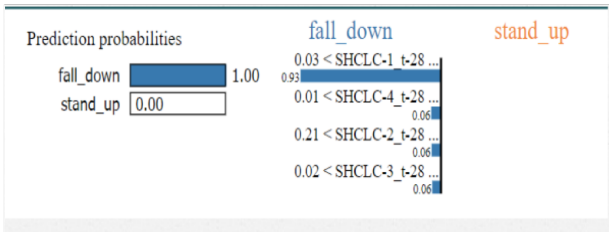


Figure 3. LIME output format for an AI-enabled fall detection system.

the central coordinate for shoulders, hips, knees, and ankles respectively. With the measurement of the Speed of Human Center Line Coordinate (SHCLC), we could identify the different kind of falls.

Figure 5 shows one kind of fall with higher speed on the movement of point 1 (i.e., SHCLC-1 on Figure 3). Therefore, the LIME outputs may provide us with the key features and reasons why the AI system judges the fall event. Also, it will indicate the specific kind of fall, such as fall over, fall down, and fall off, etc.

With domain experts' comprehensibility with the model, their trust in the AI-enabled system could be enhanced.

- User Trust

  In this research, domain experts are the key users. It is a mutual learning process during the interaction between domain experts and the ML pipeline; domain experts input their domain knowledge into the data preprocessing to confirm the training data quality for building a better model; the AI-enabled system provides insights by its data-driven analytical capabilities.

  In addition to XAI tool incorporated into the ML pipeline, domain experts provide more valuable feedback into the model refinement and training data revision through the interaction with the ML pipeline, which also help increase the user trust. An important path leading to better adoption rates identified is trust-building [18].

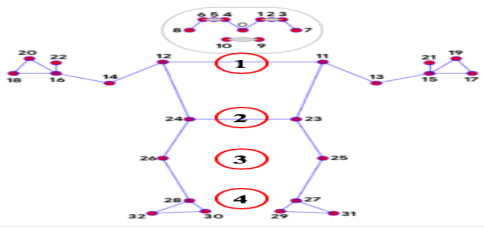


Figure 4. Human skeleton with four key coordinates on center line.

Figure 5. Different movement speed on different portion of a human while falling.

- User Adoption

  In addition to trust, Technology Acceptance Model (TAM) [19] provides us with a guideline to follow in achieving user adoption on the AI-enabled system. With TAM, the usefulness and ease of use perceived are important principles for the AI-enabled system design.

The HAC cycle starts with data engagement and then guides AI system designers and domain experts to go through the cycle to achieve a better model for deployment.

IV. Research Methodology

In this research, we design a field experiment to validate the framework of the HAC cycle. The IT artifact, hypotheses, and experiment design are described as follows:

A. IT Artifact

We select the AI-enabled fall detection system as an IT artifact, which is shown in Figure 6. There are various fall detection methods, including wearable devices with threshold setting, non-wearable device like mmWave radar and vision-based video camera. Each fall detection sensor has its advantages and disadvantages. With non-wearable sensors, people do not need to attach them on their bodies. However, they can not be used outdoors and are limited to a small area inside the detection range. However, in this research, the vision-based fall detection system is applicable for the indoor use.

B. Hypotheses

The hypotheses on domain experts' trust level are shown in Figure 7. We proposed three hypotheses(H1, H2, and H3) as follows for this research:

H1: AI-enabled system users participating in data preprocessing and model evaluation/validation with XAI incorporated would have higher trust level than users participating in data preprocessing and model evaluation/validation but without XAI incorporated.

H2: AI-enabled system users participating in data preprocessing and model evaluation/validation but without XAI incorporated would have higher trust level than users without participating in data preprocessing and model evaluation/validation, also without XAI incorporated.

H3: AI-enabled system users participating in data preprocessing and model evaluation/validation with XAI incorporated would have higher trust level than users without participating in data preprocessing and model evaluation/validation, also without XAI incorporated.

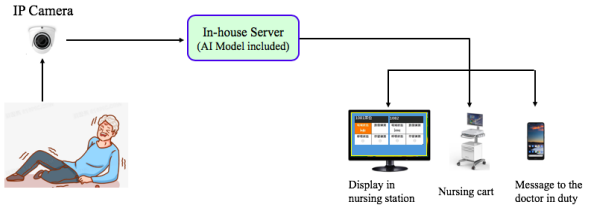


Figure 6. AI-enabled fall detection system.

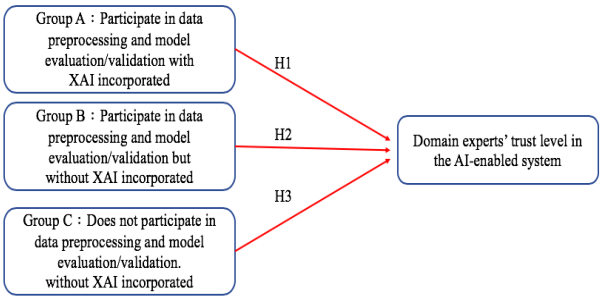


Figure 7. Hypotheses on trust level.

C. Experiment Design

More than 90 nurses, as domain experts, from one local hospital will participate in this experiment. Since user trust is one of the key components in the HAC cycle, in this research, firstly we conduct the significance level test on it.

All nurses will be divided into three groups, which are group A, B, and C. Each group will watch the same video demonstrating the brief introduction to the AI-enabled fall detection system. We designed different interaction modes with the AI-enabled system for each group, which are described as follows:

Group A: Participate in data preprocessing and model evaluation/validation with XAI incorporated.

Group B: Participate in data preprocessing and model evaluation/validation but without XAI incorporated.

Group C: As a control group, without HAC and XAI, to be informed of the system performance only.

The Likert scale will be used for trust level evaluation. The items are rated on a bipolar scale going from "I agree strongly" to "I disagree strongly", which are modified from [20]. Questions are as follows:

- I have confidence in the AI system performance.
- The AI system performance could be improved gradually.
- The output of the AI system is very predictable.
- The AI system is very reliable.
- The AI system is easy to use.
- The AI system is very efficient.
- The AI system can act as part of my team.
- I like to use the AI system.

ANOVA tool will be used for the significance analysis on trust level between groups. The timings of pre-test and post-test for each group are illustrated in Figure 8.

The nurses in Group A will be expected to have a better understanding on the key reasons of fall event identified by

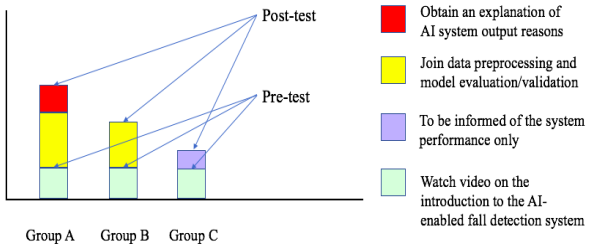


Figure 8. The timings of pre-test and post-test for each group.

the AI system. Therefore, they would have higher confidence in gradual improvement of the AI system in the development stage.

Pilot test with 9 nurses, 3 in each group, and manipulation check will be conducted to ensure the effectiveness of XAI treatment. Our basic assumption is that most nurses are rational with respect to the interpretation, provided by AI engineers, on the LIME outputs, i.e., key features.

In addition to the quantitative analysis, we will observe the differences in their interaction modes with the AI-enabled system in each group and make a complete record for qualitative analysis. For example, we have interest in the nurses' feedback or response to the XAI output explanation for one specific instance, which may encourage their engagement with the training data and test data review to assist in a better model building.

In the event that the significance level shows that nurses in Group A have the highest trust level by the introduction of XAI, the HAC cycle could be constructed with the user comprehensibility with the AI model and user adoption to the AI system. Also, a few more new test data, attributed to more data engagement, provided by the nurses in Group A would guide them to go into the second cycle for building a better model, especially in the development stage. The implementation of the HAC cycle might be considered as an approach to differentiate the user trust levels among the three groups.

V. Conclusion and Future Work

In this research, we proposed HAC Cycle based on the literature reviews, which includes four components: Data Engagement, User Comprehensibility, User Trust, and User Adoption. Also, the ML black box could be unveiled by LIME, an XAI tool, which provides the AI system users with an explanation for each instance. Hence, user trust could be built through user comprehensibility with the AI system output reasons and user adoption could be achieved under TAM.

The HAC Cycle might be considered as an approach to differentiate the user trust levels. Also, the AI model could be optimized by the implementation of this cycle with a few runs in the development stage of an AI-enabled system.

However, user comprehensibility is not limited to the user's understanding with the reasons for one specific AI system

output, which could be considered a scientific factor. User comprehensibility could also be enhanced by the model or algorithm explanation done by AI engineers. In this case, the emotional factor would be involved in the user comprehensibility. Therefore, we would propose that user comprehensibility might need to be split into two sub-components, which are AI model interpretability done by AI engineers and AI system output explainability done by XAI. The former is related to an emotional factor and the latter is related to a scientific factor. Hence, we may need both Group A1 and Group A2 to explore the differences in user trust level affected by different factors mentioned above.

ACKNOWLEDGMENT

The authors thank Alex Hsu, an AI engineer, for his assistance in the operation of XAI tool.

REFERENCES

[1] OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL 0449 G, 2022. https://www.oecd.org/digital/artificial-intelligence [retrieved: February, 2024].

[2] M. Maadi, H. A. Khorshidi, and U. Ackelin, A review on human-AI interaction in machine learning and insights for medical applications. International Journal of Environmental research and public health.Vol. 18, pp. 1-27, 2021.

[3] G. Futia and A. Vetro, On the integration of knowledge graph into deep learning models for a more comprehensible AI: Three challenges for future research. Information, Vol. 11, No. 122, pp. 1-10, 2020.

[4] D. Pedreschi et al., Meaningful explanations of black box AI decision systems. The Thirty-Third AAAI conference on Artificial Intelligence, pp. 9780-9784, 2019.

[5] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics, pp. 1-11, 2020.

[6] M. H. Jarrahi, Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Kelley School of Business, Indiana University, BUSHOR-1478, pp.1-10, 2018.

[7] I. Seeber et al., Machines as teammates: A research agenda on AI in team collaboration. Information and Management, Vol. 57, pp. 1-22, 2020.

[8] M. J. Saenz, E. R. Revilla, and C. Simon, Designing AI systems with human-machine teams. MIT Sloan Management Review, Reprint 61430, 2020.

[9] D. Wang et al., Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. Proceedings of the ACM on Human-Computer Interaction, Vol. 3, Issue CSCW, Article 211, pp. 1-24, 2019.

[10] L. Zhou, S. Paul, H. Demirkan, L. Yuan, and J. Spohrer, Intelligence augmentation: Towards building human-machine symbiotic relationship. AIS Transactions on Human-Computer Interaction, Vol. 13, Issue 2, pp. 243-264, 2021.

[11] Google Cloud, MLOps level 0: Manual process, Google Cloud Architecture Center, 2020. https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines [retrieved: February, 2024].

[12] R. S. Geiger et al., "Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data ? Quantitative Science Studies, Vol. 2, No. 2, pp. 1-42, 2021.

[13] M. T. Rebeiro, S. Signh, and C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144, 2016.

[14] A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge, Explainability and interpretability: Keys to deep medicine. Explainable AI in Healthcare and Medicine, Vol. 914, pp. 1-10, 2021.

[15] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, Trustworthy artificial intelligence: A review. ACM Computing Surveys, Vol. 55, No. 2, Article 39, pp. 1-38, 2022.

[16] M. Ghassemi, L. Oakden-Rayner, and A. Beam, The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digital Health, Vol. 3, No. 11, e745-e750, 2021.

[17] S. Jeong, S. Kang, and I. Chun, Human-skeleton based fall-detection method using LSTM for manufacturing industries. 34th International Technical Conference on Circuit/Systems, Computers and Communications, JeJu, Korea(South), pp. 1-4, 2019.

[18] P. Bedue and A. Fritzsche, Can we trust AI ? An empirical investigation of trust requirements and guide to successful AI adoption. Journal of Enterprise Information Management, Vol. 35, No. 2, pp. 530-549, 2022.

[19] F. D. Davis, A technology acceptance model for empirically testing new end-user information systems: Theory and results (Doctoral dissertation, Massachusetts Institute of Technology), pp. 1-291, 1985.

[20] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, Measuring trust in the XAI context. Technical Report, DARPA Explainable AI Program, pp. 1-26, 2018.

Can We Explain AI?: Explainable AI in the Health Domain as Told Through Three European Commission-funded Projects

Lem Ngongalah
Trilateral Research Ltd
London, UK
e-mail: Lem.ngongalah@trilateralresearch.com

Robin Renwick
Trilateral Research Ltd
Waterford, Ireland
e-mail: Robin.renwick@trilateralresearch.com

Abstract— **Artificial Intelligence (AI) has revolutionised healthcare, offering advanced diagnostics, personalised treatments, and enhanced patient outcomes. As AI increasingly integrates into healthcare systems, the need for Explainable AI (XAI) becomes paramount to ensure transparent and ethical decision-making. The lack of transparency and interpretability in AI systems poses significant challenges in healthcare, potentially undermining trust and hindering adoption. Understanding and addressing the complexities of XAI in healthcare is crucial for fostering trust among stakeholders, improving patient care, and adhering to ethical principles. Previous efforts have highlighted the importance of XAI but often lacked comprehensive approaches for implementation in diverse healthcare settings. This article explores the integration of XAI in healthcare, focusing on insights from three European Commission-funded projects under Horizon 2020/Horizon Europe. These projects prioritize transparency, accountability, and accessibility, showcasing the potential of XAI enhanced decision-making. In addition, the papers recognize the limitations of XAI, such as the absence of standardized approaches and the difficulty of balancing AI complexity with transparency, emphasizing the need for continuous refinement and adaptation to ensure successful XAI integration across varied healthcare settings.**

Keywords- ***Artificial intelligence; healthcare; explainable AI; trustworthiness; transparency***

I. INTRODUCTION

Artificial Intelligence (AI) has emerged as a transformative force in healthcare, offering unparalleled potential in diagnostics, treatment personalisation, and patient outcomes [1]. AI refers to the development of software that can use human-defined objectives to generate outputs such as content, predictions, or decisions influencing the environments they interact with [2]. AI systems encompass a broad spectrum of capabilities, with the capacity to perform varied tasks including problem-solving, learning, speech and pattern recognition, image classification and decision-making [2].

Explainable AI (XAI) is a critical component in AI design, ensuring transparent and understandable decision-making processes, in line with the AI Act [2] and Trustworthy AI guidelines [3], the foundational pillars of ethical AI development. From a computer science perspective, XAI involves developing algorithms that can provide interpretable insights into AI outcomes. In layman terms, XAI aims to provide a layer of understandability to algorithmic decision-making.

Recent controversies surrounding AI have highlighted concerns about the ethical implications of opaque decision-making processes, particularly in the healthcare sector where trust and understanding are crucial [4]. The integration of XAI is therefore not only desirable, but indispensable – imperative for the development of ethical, responsible, patient-centric, and trustworthy AI applications. However, one critical aspect of this integration is the quality, accuracy, and reliability of input data utilised by AI platforms. Healthcare data often faces challenges such as incompleteness, bias, and inconsistency, which can significantly impact performance and reliability. Addressing these data quality challenges is essential to ensure the effectiveness and trustworthiness of AI-driven decision-making processes in healthcare, thereby reinforcing the significance of XAI in healthcare settings.

This article explores the integration of XAI in healthcare, drawing insights from three European Commission-funded projects. Section I introduces the concept of Explainable AI and its significance in healthcare. Section II presents three case studies, highlighting their approaches to integrating XAI. Section III discusses the challenges and opportunities in implementing XAI in healthcare settings. Section IV concludes by outlining future directions to advance XAI in healthcare, with a focus on scalability, adaptability, and technical refinement.

II. CASE STUDIES

The iToBoS project (Grant agreement number 965221, April 2021 – March 2025) [6] aims to create an AI diagnostic platform for early skin melanoma detection, using a novel total body scanner and a computer-aided clinical decision support system that integrates patients' clinical information, genetic and imaging data, and family medical history. Despite advancements in AI, existing solutions often lack comprehensive transparency, leaving users with limited insights into decision-making processes. The project aimed to bridge this gap by integrating patient data and family medical history into a unified platform, ensuring comprehensible and transparent AI-driven diagnostics. To enhance the explainability of its AI components, the consortium prioritized transparency and interpretability through ongoing meetings within specific XAI work packages, interviews with project partners, end-users and patients, and an ongoing impact assessment process. These

assessments addressed concerns related to autonomy, transparency, and clinical effectiveness, as well as key considerations regarding patient understanding, potential conflicts between AI and clinician opinions, and the importance of clinician training. AI explainability extended to providing comprehensive insights into deep regression models. This involved adapting local XAI methods like Layer-wise Relevance Propagation for regression tasks, and using global XAI solutions, such as Concept Relevance Propagation to illustrate prediction strategies.

The COVINFORM project (Grant agreement number 101016247, November 2020 – October 2023) [7] explored the impacts of the COVID-19 pandemic across the EU member states and the UK, employing AI components to develop a risk assessment dashboard. Existing solutions often lack robustness in capturing multifaceted dimensions of vulnerability, hindering effective decision-making in pandemic response strategies. The project developed a comprehensive risk assessment dashboard, integrating statistical techniques and domain expertise to provide interpretable insights into various dimensions of vulnerability across regions and demographics, including physical, economic, social and information vulnerability. To prioritize explainability, the dashboard featured informative pop-ups/info-boxes providing interpretation guidance for dashboard outputs, and links to original data sources and metadata. End-user engagement was pivotal, employing a co-design approach involving workshops, usability testing to align technical and user needs, and cognitive walkthroughs where practitioners explored the dashboard, assessed semantic legibility, and performed tasks aligned with credible success story criteria. Recommendations from each testing phase informed subsequent usability testing rounds, refining the dashboard interface, and enhancing features relevant to the functioning and outcomes of the AI models. This iterative refinement highlights the project's commitment to user-friendly, interpretable AI-driven risk assessment tools for effective decision- making.

PREPARE-Rehab (Grant agreement number 10086219, June 2023 – May 2026) [8] aims to advance rehabilitation care for patients with chronic non-communicable diseases, by developing personalized, data-driven, computational prediction and stratification tools to enhance decision-making in selecting optimal therapy strategies. While existing solutions offer advancements in personalized medicine, they often lack transparency and user-friendliness, posing challenges in adoption and integration into clinical workflows. The project plans to address these limitations by prioritizing clear language, user-friendly interfaces, and the incorporation of graphical representations and visualization tools to enhance understanding of AI predictions. Comprehensive training for healthcare professionals is also prioritized, with emphasis on plain language and visual aids to bridge the gap between technical processes and user understanding enabling healthcare professionals to understand the advantages and limitations of AI tools. The project aims to create models with transparent decision-making processes, contributing to overall model interpretability and facilitating seamless integration of AI support in healthcare settings.

III. DISCUSSION

The three case studies presented in this extended abstract demonstrate a collective commitment to enhancing AI explainability while prioritizing transparency, accountability, and accessibility for non-technical users. By employing co-creation methodologies, these studies seek to enhance overall trustworthiness and understandability by integrating diverse perspectives throughout the development lifecycle. However, limitations exist in scaling such solutions across diverse healthcare environments, necessitating ongoing refinement and adaptation. Additionally, constraints exist in balancing the complexity of AI models with the imperative for transparency and comprehensibility, thus requiring ongoing discussion.

One significant challenge highlighted in these case studies is the lack of standardised approaches in XAI. The absence of universally accepted definitions for terms such as 'explainable' or 'interpretable' in the AI context has resulted in diverse approaches reflecting varying perspectives. This diversity complicates communication within the AI community and impedes the development of cohesive frameworks for evaluating and implementing XAI methodologies. To address this challenge, there is a need to identify and focus on specific aspects of XAI that can be standardised. By breaking down the field into identifiable parts, researchers and practitioners can work towards establishing internationally agreed standards. For instance, standardisation efforts could focus on defining key components of explainability, such as model interpretability, transparency in algorithmic decision-making, and methods for communicating AI outputs to diverse stakeholders. Furthermore, ongoing dialogue and knowledge exchange are essential for developing consensus-driven understandings of key AI concepts. Collaborative efforts, such as those within ISO [10] and CENELEC [11] play a crucial role in facilitating communication and laying the groundwork for international standards in XAI. However, uncertainties remain regarding the scalability and adaptability of co-creation methodologies in diverse cultural, industrial, and regulatory contexts. Future research and case studies, beyond the scope of EU projects, are needed to explore the broader applicability and challenges associated with scaling co-creation XAI methodologies.

IV. CONCLUSION AND FUTURE WORK

The future of XAI presents both challenges and opportunities. International standards will provide a platform for harmonised governance, conformity, and risk assessment. By identifying and standardizing key components of XAI, researchers and practitioners can facilitate smoother communication and foster sustainable innovation aligned with ethical and societal values. However, achieving this vision requires interdisciplinary collaboration, continuous dialogue, and a concerted effort to navigate evolving XAI techniques in an era of rapid technological advancement.

In our future work, we aim to further explore the scalability and adaptability of XAI methodologies, particularly across diverse cultural, industrial, and regulatory contexts. Furthermore, we plan to explore alternative approaches to addressing the varied explainability requirements for diverse stakeholders within the healthcare domain. In addition, our work will investigate the technical intricacies of implementing XAI models, including refining existing methodologies and developing novel techniques to enhance the transparency and interpretability of AI systems. Through these efforts, we aim to contribute to the ongoing evolution and refinement of XAI, ultimately enhancing trust, accountability, and accessibility in healthcare AI decision-making.

REFERENCES

[1] N. Hoppe, R. C. Härting, and A. Rahmel, "Potential Benefits of Artificial Intelligence in Healthcare," In Artificial Intelligence and Machine Learning for Healthcare, pp. 225-249, Springer, Cham, 2023.

[2] Proposal for a Regulation of The European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, Art 3(1). Accessible at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206 [retrieved: 8 February, 2024]

[3] High-Level Expert Group on AI: Ethics Guidelines for Trustworthy Artificial Intelligence. Accessible at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [retrieved: 5 February, 2024]

[4] J. P. Richardson, C. Smith, S. Curtis, S. Watson, X. Zhu, B. Barry, et al., "Patient apprehensions about the use of artificial intelligence in healthcare," NPJ digital medicine, 4(1), p.140, 2021.

[5] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, et al., "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," Information Fusion, 99, p.101805, 2023.

[6] Intelligent Total Body Scanner for Early Detection of Melanoma (ITOBOS). Accessible at: https://itobos.eu [retrieved: 5 February, 2024]

[7] COronavirus Vulnerabilities and INFOrmation dynamics Research and Modelling (COVINFORM). Accessible at: https://www.covinform.eu [retrieved: 8 February, 2024]

[8] PREPARE-Rehab. Accessible at: https://prepare-rehab.eu [retrieved: 8 February, 2024]

[9] W. J. Baumol, "The free-market innovation machine: Analyzing the growth miracle of capitalism," Princeton university press, 2002.

[10] ISO/IEC JTC 1/SC 42 - Artificial intelligence, https://www.iso.org/committee/6794475.html [retrieved: 10 February, 2024]

[11] CEN-CENELEC JTC 21 ‘Artificial Intelligence’, https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence [retrieved: 10 February, 2024]

Cancer: Investigating the Impact of the Implementation Platform on Machine Learning Models

Adedayo Seun Olowolayemo, Amina Souag, Konstantinos Sirlantzis

School of Engineering, Technology and Design, Canterbury Christ Church University (CCCU)

Canterbury, UK

email: (a.olowolayemo502, amina.souag, konstantinos.sirlantzis)@canterbury.ac.uk

***Abstract*— In the context of global cancer prevalence and the imperative need to improve diagnostic efficiency, scientists have turned to machine learning (ML) techniques to expedite diagnosis processes. Although previous research has shown promising results in developing predictive models for faster cancer diagnosis, discrepancies in outcomes have emerged, even when employing the same dataset. This study addresses a critical question: does the choice of development platform for ML models impact their performance in cancer diagnosis? Utilizing the publicly available Wisconsin Diagnostic Breast Cancer (WDBC) dataset from the University of California, Irvine (UCI) to train four ML algorithms on two distinct platforms: Python SciKit-Learn and Knime Analytics. The algorithms' performance was rigorously assessed and compared, with both platforms operating under their default configurations. The findings of this study underscore an impact of platform selection on ML model performance, emphasizing the need for thoughtful consideration when choosing a platform for predictive models' development. Such a decision bears significant implications for model efficacy and, ultimately, patient outcomes in the healthcare industry. The source code (Python and Knime) and data for this study are made fully available through a public GitHub repository.**

Keywords-Cancer; Machine Learning; Python SciKit-Learn; Knime Analytics; Wisconsin Diagnostic Breast Cancer (WDBC).

I. INTRODUCTION

Cancer is a global health menace responsible for nearly 10,000,000 deaths in year 2020 alone [1][2][3]. This disease is characterized by the uncontrolled growth of body cells which forms *tumors* classified as *malignant* - the cancerous cells that are invasive and capable of spreading to other parts of the body - or *benign* - the non-cancerous cells that are not capable of invading nearby tissues and are less harmful. This disease's complexity spans multiple organs like the breast, kidneys, brain, lungs, prostate, ovaries, and skin, posing substantial challenges for healthcare professionals and patients alike. Despite significant progress in cancer understanding and treatment development, timely diagnosis remains critical as delays exacerbate patients' conditions, often leading to irreparable outcomes and increased mortality rates.

Scientists are channeling substantial resources into accelerating the diagnostic process, and artificial intelligence, which has proven effective in various industries, is offering hope for quicker and more effective cancer diagnosis methods. Machine learning, a subset of artificial intelligence, has profoundly reshaped medical research, enhancing diagnostic precision, prognostic accuracy, and treatment strategies. By harnessing advanced computational techniques, ML algorithms ranging from Logistic Regression (LR) to Decision Trees (DT), Random Forests (RF), Gradient Boosting (GB) among several others for cancer diagnosis, extract insights from intricate medical data used in revolutionizing clinical decision-making and improving patient outcomes from pinpointing diseases through image analysis [4] to forecasting patient responses to therapies [5].

These ML algorithms have showcased remarkable potential in the field. However, a critical aspect that we found to be underexplored is the impact of implementation platforms on which the algorithms are trained, and models are developed, such as Python Scikit-learn and Knime analytics, on the performance of these algorithms. Therefore, understanding the nuanced influence of implementation platforms on ML algorithms is pivotal.

Against this backdrop, this study used supervised learning, training models on labeled WDBC datasets [6] to evaluate the performance metrics of ML algorithms including accuracy, precision, recall, and F1-Score focusing on the nuanced relationship between implementation platforms and the efficacy of these algorithms. It emphasizes the potential impact of platform choice on algorithm behavior, highlighting the necessity of discerning these disparities.

This study embarks on two pivotal inquiries:

(1) It seeks to answer whether the choice of the implementation platform significantly impacts the performance of ML algorithms in cancer data classification,

(2) Identifies which of the selected algorithms performed best in cancer dataset binary classification task.

By delving into these fundamental questions and meticulously avoiding hyperparameter tuning, this research provides nuanced insights, offering a comprehensive understanding of the intricate interplay between ML

algorithms, implementation platforms, and feature significance.

The rest of this paper is organized as follows: Works relating to this study were explored in Section II, examining relevant literature to the research question. The section starts by looking at studies that used ML in cancer research, then at the different algorithms implemented, the train-test split, performance metrics, dataset sources, and implementation platforms used. Section III outlines the Methodology used for this study, detailing data collection and pre-processing steps, feature selection, and implementation of the selected ML models. Section IV presents the Results and Discussion, followed by Conclusion and Future Work in Section V.

II. RELATED WORK

Researchers have explored and reported the use of various supervised ML algorithms in different areas of human health and medical fields. Some previous studies reviewed are briefly discussed below.

A. Machine Learning in Cancer Research

Michael et al. in [7] tested five ML classification algorithms on 912 breast ultrasound images found that Light Gradient Boosting Machine (LightGBM), the algorithm proposed in their work, which has an accuracy of 99.86%, outperformed other algorithms including the K-Nearest Neighbour (KNN), and RF in binary classification of cancerous cells as either malignant or benign. Similarly, Ara et al. in [8] used a ML techniques to develop model for classifying cancer cells into two main categories. Kumar et al. in [9] on the other hand focused on using ML ensemble techniques for breast cancer detection and classification. Their Optimized Stacking Ensemble Learning (OSEL) model showed a higher accuracy in performing the task than other ensemble ML techniques, such as Stochastic Gradient Boosting and XGBoost tested in their research. Ebrahim et al. in [10] tested eight predictive algorithms on National Cancer Institute dataset to identify which algorithm would predict cancer cell more accurately.

B. Selection of Algorithm

LR, a linear model is a powerful predictive analysis tool that is especially useful for binary classification [11]. Rahman et al. [12] examined six ML algorithms for predicting Chronic Liver Disease (CLD) and LR algorithm was found to be the most effective in predicting CLD based on the selected features. Zhu et al. in [11] experimented with improved LR in the classification of binary variable and one or more independent variables to predict diabetes.

Likewise, Tree based algorithms including DT, RF and GB are widely researched with the intent of harnessing their strengths particularly in performing classification tasks. DT serve as foundational structures, offering transparency and interpretability by partitioning feature spaces into hierarchical branches thereby excelling in capturing non-linear relationships and feature interactions, enabling straightforward visualization of decision-making processes. Moving beyond individual trees, RF combines multiple DT through ensemble techniques, averting overfitting and increasing predictive accuracy [13]. By combining varied perspectives from individual trees, RF provides robust generalization and robustness to noisy data.

By extension, GB algorithm, a more advanced method, embraces an iterative refinement to enhance predictive performance and in particular, Gradient Boosting Trees, such as XGBoost. It employs sequential tree fitting to target the residuals of prior iterations, systematically improving model predictions. These algorithms perform better in modeling complex relationships, accommodating non-linearities, and excelling in predictive accuracy across domains [14][15]. These characteristics formed the basis on which we selected the algorithms in our study.

C. Train-Test Split

For evaluation, datasets used in various studies are split into different proportions using the larger proportion to train algorithms while the smaller proportion is used to test at the inference stage of model development. In [10], the authors assessed the performance of some classical and deep learning algorithms used to predict breast cancer, including DT, LR, KNN, Support Vector Machine (SVM), Recurrent Neural Networks (RNN) and Ensemble Learning. They used Train/Test split of 70:30 and 90:10. DT and Ensemble methods showed higher accuracy both before and after feature selection. Whereas DT did not perform optimally in Kidney Cancer Lung Metastasis prediction as reported by [16] when trained with 52,222 data from Surveillance, Epidemiology, and End Results (SEER) database and 492 hospital patient data with Train/Test split of 70:30 returning accuracy of 82% which is significantly lower than in other studies reviewed.

D. Performance Metrics

Efficient model development and deployment require rigorous assessment, evaluating the accuracy and other key metrics like precision, recall, and F1-score derived from the confusion matrix. Accuracy gauges correctly predicted instances against the total dataset, offering a general overview of predictive success. In imbalanced datasets, relying solely on accuracy can be deceptive. Therefore, other metrics such as precision, recall, and F1-score gain importance. Precision specifically gauges correctly predicted positive instances, which is crucial in scenarios like medical diagnoses where false positives can have adverse consequences. Recall assesses true positive predictions, essential for capturing all positive instances, especially critical in medical scenarios to avoid missing dangerous conditions. F1-score strikes a balance between precision and recall, offering a nuanced evaluation, particularly valuable when dealing with class imbalances in datasets.

These four metrics were assessed in our study (**TABLE 3**); they collectively provide a comprehensive assessment of a model's performance.

E. Datasets

Data quality is fundamental in machine learning, shaping model development and real-world utility. The WDBC [6]

has been pivotal in healthcare, especially for binary tumor classification, crucial in timely cancer detection and treatment planning. While studies like [17][18][19] employed smaller, open-source WDBC datasets (typically fewer than 600 records and 30 features), other studies in [10] and [15] diverged. For example, [10] used a substantial dataset from the National Cancer Institute (NIH) containing 1.7 million records and 210 features. Despite its size, dataset quality, marked by precision and representativeness, significantly influences outcomes. Smaller datasets with these qualities outperform larger, noisier ones. This distinction is evident in accuracy rates, with open-source datasets achieving 99.12%, 99.67%, and 100%, compared to the model in [10] with a lower accuracy of 98.7%.

F. Implementation Platform

KNIME Analytics, a no-code tool recognized for its user-friendly interface and compatibility with various other tools, has been utilized for comprehensive ML research, as demonstrated in studies like [20] which looks at cancer incidence among individuals with HIV in Zimbabwe. Meanwhile, Python, with its extensive ecosystem and libraries like SciKit-Learn, has gained prominence in machine learning. Studies in [16][21][22] performed their cancer research work using Python. Both platforms have strong support from scientists, underlining the need for further research into their respective impacts on algorithm performance.

The findings of the literature are summarized in **TABLE 1.** The table highlights the latest studies that used ML techniques in cancer research, the data source used, train – test split ratio adopted in the study, the implementation platform used, the algorithm type and the model accuracy (a '–' has been used in the table in the case where the information was missing in the literature).

The recent surge in research on ML applications in healthcare, specifically in diverse cancer data sets, is evident. Nevertheless, a significant research gap persists concerning the impact of implementation platforms on algorithm performance in cancer classification.

While several studies have used different implementation platforms in developing ML models for predictive and classification tasks, none, to the best of our knowledge, have examined the impact of implementation platforms on ML algorithm performance. This gap forms the focal point of our research contribution, that will be explored in subsequent sections, highlighting the novelty and importance of our investigation.

TABLE 1. COMPARATIVE REVIEW OF SOME STUDIES THAT USED MACHINE LEARNING TECHNIQUES IN CANCER RESEARCH.

Author, Year	Data Source	No of Records /Features	Train/Test Split	Implementation Platform	Algorithm Type	Model Accuracy
Ebrahim et al. [10],2023	National Cancer Institute (NIH), USA	70,079/107	70:30 &90:10	Python	DT, LR, VM, LD, ET, KNN	98.7%
Shafique et al.[18],2023	Kaggle	569/30	75:25	-	RF, VM, GBM, LR, MLP, KNN	100%
Uddin et al. [19], 2023	UCI	569/30	70:30	Python	SVM, RF, KNN, NB, DT, LR, AB, GB, MLP, NCC, VC	98.7%
Zhang et al [23]., 2022	TCGA	604/ -	-	R & Python	RF, SVM, libD3C	99.67%
Aamir et.al.[24], 2022	UCI	569/26	80:20 &70:30	Python & Tensor Flow	RF, GB, SVM, ANN, MLP	99.12%
Yi et al., [16],2023	SEER& Southwest Hospital, China.	52,714 / -	70:30	Python	LR, XGB, RF, SVM, ANN, DT	-
Mahesh et al., [22],2022	Kaggle	143/10	70:30	Python	NB, AltDT, RedEPT, RF	98.20%
ATEŞ et al. [25] 2021	Kaggle	569/30	70:30	Knime	NB, DT, MLP	96.5%
Minnoor et al.[13] 2023	UCI	569/30	80:20	-	RF, SVM, DT, MLP, KNN	100%
Ara et al. [8] , 2021	UCI	569/30	75:25	-	SVM, LR, KNN, DT, NB, RF	96.5%
Liu, et al. [26]2018	UCI	569/30	75:25	Python	LR	96.5%

Addressing the gap identified in the literature, the next section presents the methodology carried out.

III. METHODOLOGY

This study's methodology comprises systematic steps for a comparative analysis of ML algorithms using the WDBC dataset and two implementation platforms. The process as illustrated in includes data collection, exploration, feature engineering, and selection using filtering and random forest techniques. The dataset was split into an 80% training set and a 20% test set before model development, ensuring a robust evaluation process.

A. Data Collection and Preprocessing

We selected a publicly available dataset on UCI Machine Learning repository, the WDBC [6] because it was sourced from a medical research study and its extensive use in breast cancer machine learning research due to its real-world applicability, in addition to its popularity within the research community for binary classification task. With 569 occurrences and 30 attributes (benign tumours made up 62.7% of the total instances while the cancerous tumour, malignant class comprise 37.3%) was extracted from digitized Breast Mass Fine Needle Aspiration (FNA)

specimens, including features like "Diagnosis" (categorized as Malignant (M) or Benign (B)) and various measurements from cell nuclei in biopsy images ("radius_mean," "texture_mean," "perimeter_mean," etc.) [6], providing a rich foundation for cancer predictive analysis.

TABLE 2. WDBC DATASET VARIABLES DATATYPE.

```
Data columns (total 32 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   id                      569 non-null    int64
 1   diagnosis               569 non-null    int32
 2   radius_mean             569 non-null    float64
 3   texture_mean            569 non-null    float64
 4   perimeter_mean          569 non-null    float64
 5   area_mean               569 non-null    float64
 6   smoothness_mean         569 non-null    float64
 7   compactness_mean        569 non-null    float64
 8   concavity_mean          569 non-null    float64
 9   concave points_mean     569 non-null    float64
 10  symmetry_mean           569 non-null    float64
 11  fractal_dimension_mean  569 non-null    float64
 12  radius_se               569 non-null    float64
 13  texture_se              569 non-null    float64
 14  perimeter_se            569 non-null    float64
 15  area_se                 569 non-null    float64
 16  smoothness_se           569 non-null    float64
 17  compactness_se          569 non-null    float64
 18  concavity_se            569 non-null    float64
 19  concave points_se       569 non-null    float64
 20  symmetry_se             569 non-null    float64
 21  fractal_dimension_se    569 non-null    float64
 22  radius_worst            569 non-null    float64
 23  texture_worst           569 non-null    float64
 24  perimeter_worst         569 non-null    float64
 25  area_worst              569 non-null    float64
 26  smoothness_worst        569 non-null    float64
 27  compactness_worst       569 non-null    float64
 28  concavity_worst         569 non-null    float64
 29  concave points_worst    569 non-null    float64
 30  symmetry_worst          569 non-null    float64
 31  fractal_dimension_worst 569 non-null    float64
dtypes: float64(30), int32(1), int64(1)
```

In the data preprocessing phase, the dataset was structured into a Python dataframe named "breast". The data was subsequently queried to ascertain the data types and to check for presence of any null values. According to TABLE 2, data consists of both integer and floating-point values, and no null values were found. Further analysis involved identifying outliers through box plots and the Capping method was applied to mitigate their impact. This technique, as presented by [27] involved setting values below the lower whisker to the lower whisker's value and values above the upper whisker to the upper whisker's value, ensuring an unbiased model.

Normalization was achieved through Z-Score Normalization (Standardization) which rescales each feature to normal distribution with a mean of 0 and a standard deviation of 1 [28][29]. Standardizing features to the same scale are essential to prevent algorithms from giving undue importance to larger-magnitude features, thus preserving fairness and accuracy across diverse ML algorithms. This process ensured that each feature contributed proportionally to the learning process, averting dominance by any single feature, and promoting balanced model decisions. Equation 1 below represents the computation formula for z-score standardization [29].

$$Z=(x-\mu)/\sigma. \quad (1)$$

where z is the scaled value of the feature,
x is the original value of the feature,
μ is the mean value of the feature, and
σ is the standard deviation of the feature.

Correlation analysis was conducted to evaluate the relationship between each feature, a crucial step preceding feature selection, providing insights into features independently related to the target variable. This analysis was followed by a detailed examination of individual feature relationships, discerning the impact of changes in one feature on another and identifying strongly correlated independent features. High correlation between features suggests redundancy, potentially diminishing their value in the model, thus ensuring more effective predictions.

B. Feature Selection

Selection of essential features is a crucial stage [30]. We employed both the Filter Method as in [4], and the Tree-Based Method as in [31]. Initially, the Filter Method was utilized to evaluate dataset features based on their correlation scores with the target variable. Features with coefficients $\leqslant 0.5$ were eliminated as they were considered to have low significance based on feature selection technique used in [32], while those above this threshold were retained, resulting in the identification of 15 out of the 30 features for further analysis. To confirm these selections, the Tree-Based Method was employed, utilizing the RF Classifier. This method, known for balancing interpretability and computational efficiency while capturing both linear and non-linear relationships between the features as shown in **Figure 2**, affirmed the chosen features, underscoring their significance in model development [30].

The synergy between the two methods ensured a comprehensive and accurate feature selection process, crucial for enhancing the model's predictive capabilities.

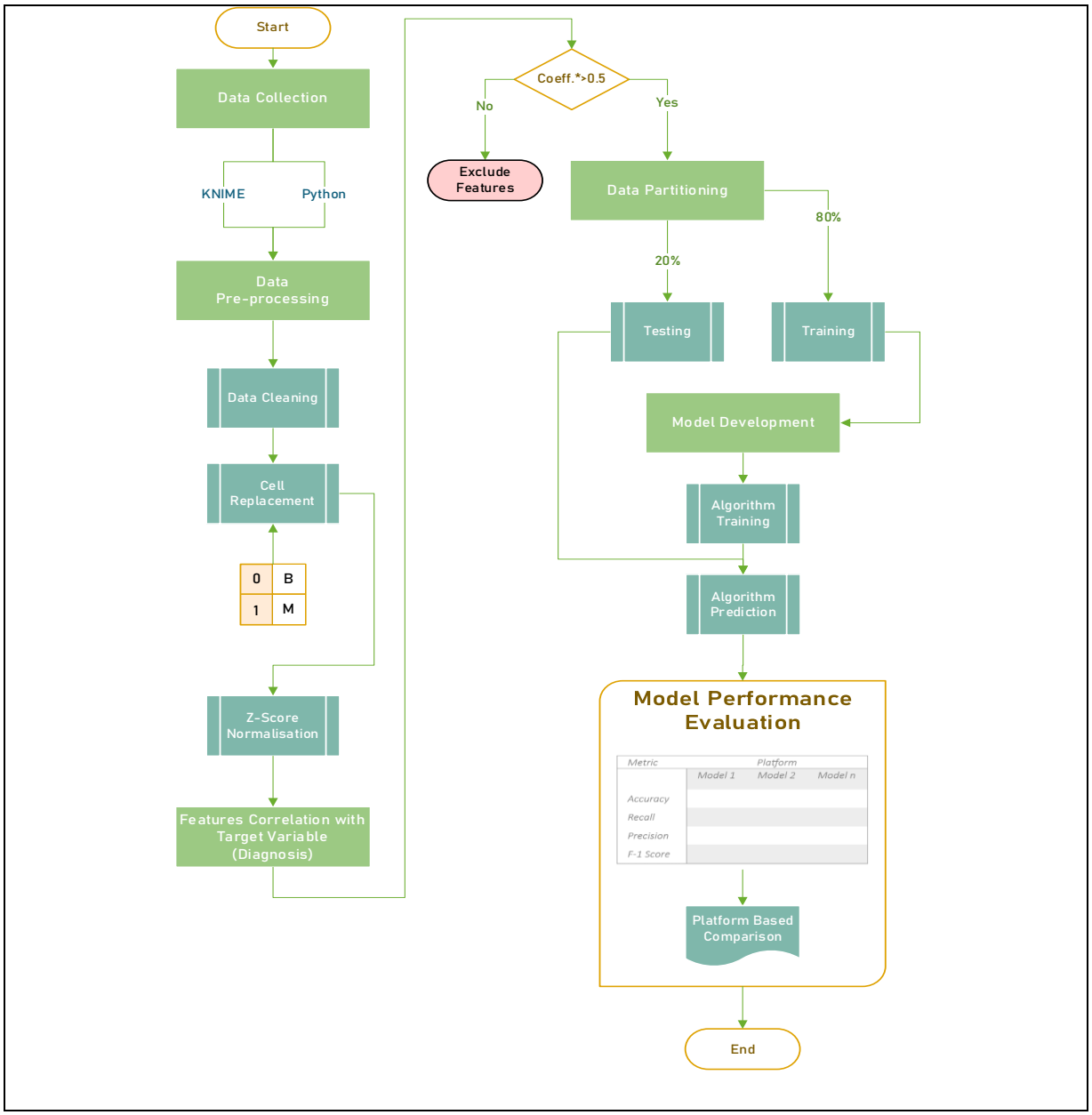


Figure 1. Flowchart illustrating the research methodology applied in this study.

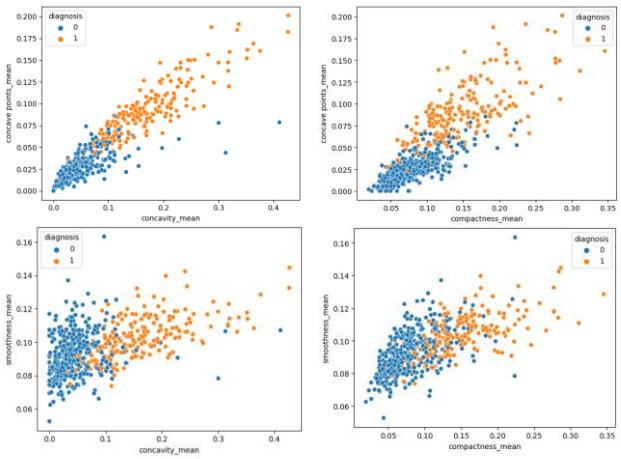


Figure 2. Scatter plot showing relationships between some of the features. (A view of relationships between some other features can be viewed on the GitHub [31]).

Understanding the relationship between the features helped to inform the class of ML algorithms that will be best suited for the classification task.

C Model Selection and Implementation

Four Supervised ML classification algorithms were chosen, each based on their specific properties and extensive use in previous research. This study selected LR because of its ability to estimate outcome probabilities, along with its interpretability and computational efficiency. These attributes make LR a widely favoured option for binary classification tasks. DT, RF, and GB, all belonging to the Tree-Based algorithms category, were selected for their recursive partitioning approach, which efficiently identifies optimal features and split points, enhancing the models' accuracy.

This study was carried out utilizing the Knime Analytics Platform Version 4.7.6 and Python version 3.11.4 (Jupyterlab) using the Scikit-Learn library. During this process, the algorithms underwent training and testing in their default configurations, with a maximum of 100 epochs, a learning rate of 0.1, and no parameter tuning—except in Knime, where the default split criterion for RF was adjusted from "Information Gain Ratio" to "Gini Index," aligning it with the default split criterion in Scikit Learn.

This adjustment was implemented to maintain fairness in the comparative evaluation. A train-test split ratio of 80:20 was applied, with 80% of the dataset allocated for training, enabling the algorithms to learn patterns, while the remaining 20% was reserved for testing, evaluating the models' ability to generalize to unseen data points. This methodology ensured a comprehensive evaluation of the algorithms' performance and their suitability for the classification task at hand. The source code (Python and Knime) and data for this study can be found in the public GitHub repository [31].

IV. Results and Discussion

This section outlines the experimental results achieved following implementation of the four algorithms on both platforms comparatively in **TABLE 3** and visualized in **Figure 3** after assessing their Accuracy, Recall, Precision, and F1-Score.

TABLE 3. COMPARATIVE ASSESSMENT OF MODEL PERFORMANCE ON THE TWO PLATFORMS.

Algorithm	Tool	Accuracy	Recall	Precision	F1-Score
LR	SciKit-Learn	0.956	0.929	0.951	0.940
	Knime	0.921	0.884	0.905	0.894
DT	SciKit-Learn	0.930	0.952	0.870	0.909
	Knime	0.886	0.907	0.813	0.857
RF	SciKit-Learn	0.947	0.976	0.891	0.932
	Knime	0.912	0.884	0.884	0.884
GB	SciKit-Learn	0.974	0.976	0.953	0.965
	Knime	0.904	0.861	0.881	0.871

Also, we reported the Confusion matrix, showing the True Positive, True Negative, False Positive, and False Negative values, providing a comprehensive evaluation of this study's outcomes in **TABLE 4**.

TABLE 4. PLATFORM BASED CONFUSION MATRIX OF THE ALGORITHMS.

SciKit-Learn

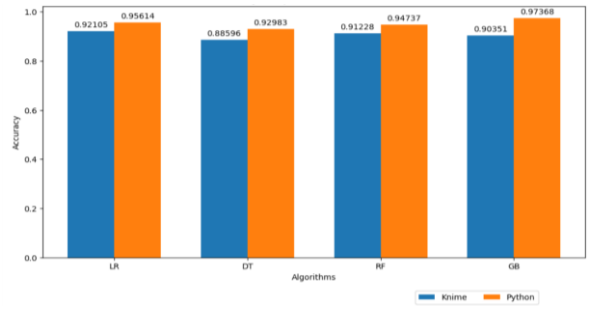
	Logistic Regression		Decision Tree	
	Negative	Positive	Negative	Positive
Negative	70	2	66	6
Positive	3	39	2	40

	Random Forest		Gradient Boosting	
	Negative	Positive	Negative	Positive
Negative	67	5	70	2
Positive	1	41	1	41

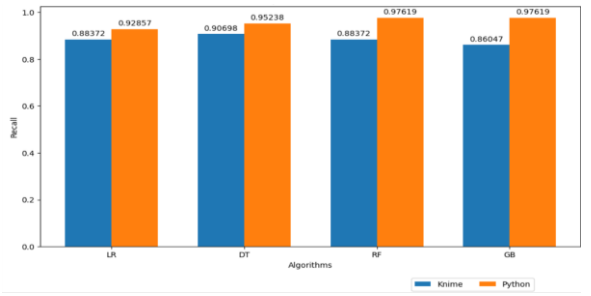
Knime

	Logistic Regression		Decision Tree	
	Negative	Positive	Negative	Positive
Negative	67	4	62	9
Positive	5	38	4	39

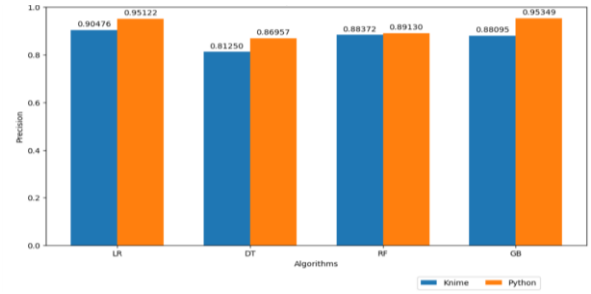
	Random Forest		Gradient Boosting	
	Negative	Positive	Negative	Positive
Negative	66	5	66	5
Positive	5	38	6	37



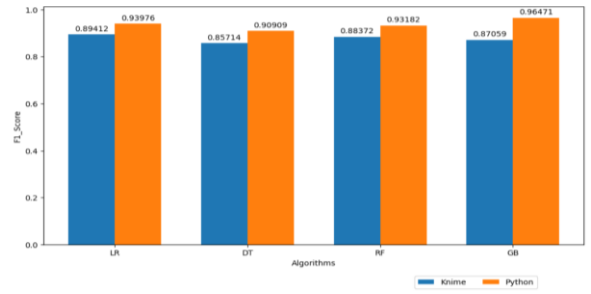
(a)



(b)



(c)



(d)

Figure 3. Column Chart Visualization-Comparison of all the algorithms performance on both platforms for:
(a) Accuracy (b) Recall (c) Precision and (d) F1-Score.

In the KNIME Analytics platform, the LR algorithm achieved an Accuracy of 0.92105, with Recall, Precision, and F1 Score of 0.88372, 0.90476, and 0.89412, respectively signifying that the model correctly classified approximately 92.11% of the instances. In comparison, the DT algorithm demonstrated a slightly lower Accuracy of 0.88596, yet it exhibited higher Recall (0.90698) and F1 Score (0.85714), suggesting that it is proficient in capturing true positive instances while maintaining a balance between precision and recall, although its Precision score was 0.81250, indicating a relatively lower ability to avoid false positives.

The RF algorithm on the other hand achieved an Accuracy of 0.91228, almost on par with LR. It yielded Recall, Precision, and F1 Score of 0.88372, 0.88372, and 0.88372, respectively, presenting consistent performance across the metrics. The GB algorithm, like DT, secured an Accuracy of 0.90351, while it demonstrated a Recall of 0.86047, Precision of 0.88095, and F1 Score of 0.87059 reflecting a balanced trade-off between sensitivity and precision, critical in medical diagnosis scenarios.

However, on Python (Scikit-Learn) platform, the LR model exhibited superior performance, with an Accuracy of 0.95614. This shows an improvement in predictive accuracy when compared to its counterpart in KNIME Analytics. The Recall 0.92857, Precision 0.95122, and F1 Score 0.93976 further validate the model's proficiency in correctly classifying instances. DT and RF algorithms also displayed an improvement in their performance in the Python (Scikit-Learn) environment, with Accuracy values of 0.92981 and 0.94737, respectively.

Moreover, the Recall, Precision, and F1 Score values for these models witnessed an increase, thereby strengthening their overall predictive capabilities. The GB shows remarkable performance, attaining an Accuracy of 0.97368, a significant improvement compared to its counterpart in KNIME corroborating its impressive performance with Recall, Precision, and F1 Score values of 0.97619, 0.95349, and 0.96471, respectively, making it a standout in terms of all metrics.

The comparative analysis of these algorithms across the two platforms demonstrates the intricate relationship between algorithm choice, implementation environment, and resultant performance metrics. While KNIME Analytics rendered reliable results, Python (Scikit-Learn) emerged as the platform offering enhanced predictive accuracy across the board. Notably, the GB algorithm stood out in Python (Scikit-Learn), exhibiting remarkable performance, which is highly relevant in medical contexts where accurate classification holds paramount importance. These findings underscore the necessity of carefully considering both algorithm selection and platform for optimal performance in predictive modeling endeavors.

Additionally, the confusion matrix of the models was evaluated on their ability to predict both the 'Positive' and 'Negative' classes, and the calculated metrics offer valuable insights into their proficiency. The matrices revealed that while models generally perform well, some algorithms, such as DT and GB, consistently exhibit a higher number of True Positives emphasizing the accurate prediction of positive cases which is crucial in medical contexts to minimize the risk of false negatives.

Comparing the KNIME and SciKit-Learn platforms, a pattern emerges. Generally, the SciKit-Learn platform showcases slightly better performance metrics, particularly in terms of True Positives and True Negatives. This disparity suggests that the SciKit-Learn implementation may have certain advantages in terms of predictive accuracy and class separation.

Also in our analysis, we conducted a comparative assessment of the LR, GB, and RF models on scikit-learn against the Baseline Model Performance (BMP- available on UCI website) established using the same dataset from the UCI Machine Learning Repository. The LR and GB models demonstrated accuracy values of 95.6 and 97.4, respectively, falling within the BMP range [92.308-98.601]. Similarly, their precision scores (95.1 and 95.3) were consistent with the baseline range [91.555-98.576]. In contrast, the RF model reported accuracy and precision scores (94.7 and 89.1) below the lower limit of the baseline performance. On the other hand, for all metrics, the performances of the algorithms on Knime Analytics were lower than the lower limits of the BMP score.

V. Conclusion and Future Work

This comparative experiment aimed to investigate the potential impact of machine learning implementation platform on the performance of machine learning models using the WDBC dataset and four classification algorithms during both training and inference phases in Python SciKit-Learn and Knime Analytics. The results demonstrated variation in the metrics for the algorithms in Python compared to Knime. While Knime showed its strength with the LR algorithm in terms of accuracy, Python presented different performance patterns, with DT excelling in recall and RF as well as GB providing high recall values, which are crucial in the context of cancer diagnosis as it suggests a reduced likelihood of false negatives.

These findings emphasize the significance of platform choice when considering the specific performance metrics required for a given application, shedding light on the intricate relationship between algorithm selection and the implementation environment. It is important to note that this study does not intend to render a verdict on the overall efficacy of either tool in ML model development but rather serves as an investigation into the potential disparities introduced by their respective architectures, providing insights for informed decision-making in predictive modeling endeavors.

Further research should explore a larger dataset as we hope this may contribute to the generalizability of the models and as a means of broadening the applicability of these findings. Future studies may also evaluate the performance of the algorithms on both platforms using other datasets. In addition, future work may:

(i) drill down to identify factors responsible for the observed differences by examining the platforms architecture,

(ii) extend the experiment by including some other classifiers algorithms, such as SVM and Multi-Layer Perceptron (MLP).

(iii) implement on different platforms including R and Weka or test multiple datasets.

References

[1] B. S. Chhikara and K. Parang, “Global Cancer Statistics 2022: the trends projection analysis” *Chemical Biology Letters*, vol. 10, no. 1, Article no. 1, January. 2023, doi: https://scholar.google.com/scholar?q=urn:nbn:sciencein.cbl.2023.v10.451

[2] “*CANCER FACT SHEETS* - Global Cancer Observatory.” Accessed: Feb. 07, 2024. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf

[3] V. D. P. Jasti et al., “Computational Technique Based on Machine Learning and Image Processing for Medical Image Analysis of Breast Cancer Diagnosis” *Security and Communication Networks*, vol. 2022, p.1-7, March. 2022, doi: 10.1155/2022/1918379.

[4] J. Kong et al., “Network-based machine learning approach to predict immunotherapy response in cancer patients” *Nature communications*, vol. 13, no. 1, Article no. 1, June 2022, doi: 10.1038/s41467-022-31535-6.

[5] W. Wolberg, O. Mangasarian, and W. Street, “Breast Cancer Wisconsin (Diagnostic).” UCI Machine Learning Repository, 1995. doi: 10.24432/C5DW2B.

[6] E. Michael, H. Ma, H. Li, and S. Qi, “An Optimized Framework for Breast Cancer Classification Using Machine Learning” BioMed Research International, vol. 2022, p. e8482022, Feb. 2022, doi: 10.1155/2022/8482022.

[7] S. Ara, A. Das, and A. Dey, “Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms” in 2021 International Conference on Artificial Intelligence (ICAI), Apr. 2021, pp. 97–101. doi: 10.1109/ICAI52203.2021.9445249.

[8] M. Kumar, S. Singhal, S. Shekhar, B. Sharma, and G. Srivastava, “Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning” *Sustainability*, vol. 14, no. 21, Aricle. no. 21, January. 2022, doi: 10.3390/su142113998.

[9] M. Ebrahim, A. A. H. Sedky, and S. Mesbah, “Accuracy Assessment of Machine Learning Algorithms Used to Predict Breast Cancer” *Data*, vol. 8, no. 2, Article no. 2, Feb. 2023, doi: 10.3390/data8020035.

[10] C. Zhu, C. U. Idemudia, and W. Feng, “Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques” *Informatics in Medicine Unlocked*, vol. 17, p. 100179, January 2019, doi: 10.1016/j.imu.2019.100179.

[11] A. K. M. Rahman, F. M. Shamrat, Z. Tasnim, J. Roy, and S. Hossain, “A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms”. International Journal of Scientific & Technology Research. vol. 8, pp. 419–422, Nov. 2019.

[12] M. Minnoor and V. Baths, “Diagnosis of Breast Cancer Using Random Forests” *Procedia Computer Science*, vol. 218, pp. 429–437, Jan. 2023, doi: 10.1016/j.procs.2023.01.025.

[13] W. Li, Y. Yin, X. Quan, and H. Zhang, "Gene Expression Value Prediction Based on XGBoost Algorithm" *Frontiers in Genetics*, vol. 10, 2019, Accessed: Feb. 07, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fgene.2019.01077

[14] X. Wan, "Influence of feature scaling on convergence of gradient iterative algorithm". *Journal of physics: Conference series*., vol. 1213, no. 3, p. 032021, Jun. 2019, doi: 10.1088/1742-6596/1213/3/032021.

[15] X. Yi et al., "Development and External Validation of Machine Learning-Based Models for Predicting Lung Metastasis in Kidney Cancer: A Large Population-Based Study" *International Journal of Clinical Practice*, vol. 2023, p. e8001899, Jun. 2023, doi: 10.1155/2023/8001899.

[16] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine Learning Classification Techniques for Breast Cancer Diagnosis" *IOP Conference Series: Materials Science and Engineering*, vol. 495, no. 1, p. 012033, Apr. 2019, doi: 10.1088/1757-899X/495/1/012033.

[17] R. Shafique et al., "Breast Cancer Prediction Using Fine Needle Aspiration Features and Upsampling with Supervised Machine Learning" *Cancers*, vol. 15, no. 3, Art. no. 3, Jan. 2023, doi: 10.3390/cancers15030681.

[18] K. M. M. Uddin, N. Biswas, S. T. Rikta, and S. K. Dey, "Machine learning-based diagnosis of breast cancer utilizing feature optimization technique" *Computer Methods and Programs in Biomedicine Update*, vol. 3, p. 100098, Jan. 2023, doi: 10.1016/j.cmpbup.2023.100098.

[19] T. Shamu et al., "Cancer incidence among people living with HIV in Zimbabwe: A record linkage study" *Cancer Reports*, vol. 5, no. 10, p. e1597, 2022, doi: 10.1002/cnr2.1597.

[20] Q. T. N. Nguyen et al., "Machine learning approaches for predicting 5-year breast cancer survival: A multicenter study" *Cancer Science*, vol. 114, no. 10, pp. 4063–4072, 2023, doi: 10.1111/cas.15917.

[21] T. R. Mahesh, V. Vinoth Kumar, V. Muthukumaran, H. K. Shashikala, B. Swapna, and S. Guluwadi, "Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast Cancer" *Journal of Sensors*, vol. 2022, p. e4649510, Sep. 2022, doi: 10.1155/2022/4649510.

[22] Y. Zhang et al., "Machine learning-based prognostic and metastasis models of kidney cancer" *Cancer Innovation*, vol. 1, no. 2, pp. 124–134, 2022, doi: 10.1002/cai2.22.

[23] S. Aamir et al., "Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques" *Computational and Mathematical Methods in Medicine*, vol. 2022, p. e5869529, Aug. 2022, doi: 10.1155/2022/5869529.

[24] İ. Ateş and T. T. Bilgin, "The Investigation of the Success of Different Machine Learning Methods in Breast Cancer Diagnosis" *Konuralp Medical Journal*, vol. 13, no. 2, Article. no. 2, Jun. 2021, doi: 10.18521/ktd.912462.

[25] L. Liu, "Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning" in 2018 International Conference on Robots & Intelligent System (ICRIS), May 2018, pp. 157–160. doi: 10.1109/ICRIS.2018.00049

[26] X. Feng, Y. Cai, and R. Xin, "Optimizing diabetes classification with a machine learning-based framework" *BMC Bioinformatics*, vol. 24, no. 1, p. 428, Nov. 2023, doi: 10.1186/s12859-023-05467-x.

[27] S. Sumin, "The Impact of Z-Score Transformation Scaling on the Validity, Reliability, and Measurement Error of Instrument SATS-36," JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia), vol. 11, no. 2, Art. no. 2, Nov. 2022.

[28] M. Pagan, M. Zarlis, and A. Candra, "Investigating the impact of data scaling on the k-nearest neighbor algorithm" *Computer Science and Information Technologies*, vol. 4, no. 2, Art. no. 2, Jul. 2023, doi: 10.11591/csit.v4i2.p135-142.

[29] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective" *Neurocomputing*, vol. 300, pp. 70–79, Jul. 2018, doi: 10.1016/j.neucom.2017.11.077.

[30] G. Alfian et al., "Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method" *Computers*, vol. 11, no. 9, Art. no. 9, Sep. 2022, doi: 10.3390/computers11090136.

[31] A.Olowolayemo (2023), Cancer3IPMLM, GitHub: https://github.com/ProfDee92/Cancer-3IPMLM/blob/main/README.md

[32] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction" *Frontiers in Bioinformatics*, vol. 2, p. 927312, 2022.

Predictive Analytics for Emergency Department Visits Based on Local Short-Term Pollution and Weather Exposure

Isabella Della Torre
R&D Researcher
GPI SpA
Trento, Italy
email: isabella.dellatorre@gpi.it

Ismaela Avellino
R&D Researcher
GPI SpA
Trento, Italy
email: ismaela.avellino@gpi.it

Francesca Marinaro
R&D Researcher
GPI SpA
Trento, Italy
email: francesca.marinaro@gpi.it

Andrea Buccoliero
R&D Project Manager
GPI SpA
Trento, Italy
email: andrea.buccoliero@gpi.it

Antonio Colangelo
R&D Director
GPI SpA
Trento, Italy
email: antonio.colangelo@gpi.it

***Abstract*-Proper management of Emergency Rooms is needed to improve healthcare and patient satisfaction. Predicting accesses and hospitalisation rates through Machine Learning approaches appears promising, especially when coupled with air pollution and weather data. This work applies both Random Forest and AutoRegressive Integrated Moving Average approaches on data related to Brescia's clinical and environmental data from 2018 to 2022 to predict daily accesses or daily hospitalisations for cardiovascular and respiratory disorders. The predictions adhere quite well to the actual data for Random Forest, but less for AutoRegressive Integrated Moving Average. However, even if the specific value is not always correctly predicted, the overall trend seems to be rightly forecasted and performance metrics are mostly satisfying. Although additional work is required to improve their performances, results are encouraging and this sort of geographically-localised time-series forecasting seems feasible. Future developments will take into consideration the whole province of Brescia.**

***Keywords-Forecasting; ER accesses; Hospitalisation; Pollution; Weather*.**

I. INTRODUCTION

Being able to properly manage the Emergency Department (ED) and Emergency Room (ER) is crucial to provide functional healthcare and improve patients' satisfaction [1]. This leads to a strong need for accurately predicting visitor volume and patient admissions to facilitate the planning of resources and staff for the whole hospital.

Multiple researchers have tried to predict access and admission rates based on historical ED data by creating scores or using deep learning (DL) or machine learning (ML) models (like Recurrent Neural Networks, Logistic Regression, Random Forest or Extreme Gradient Boosting) to forecast daily accesses to the ER [2] [3], the possibility of a patient's hospital admission after going through the triage [4] or even the risk of death [5]. Results were so encouraging, that others looked for associations with the surrounding environment.

In fact, there is proof that weather affects one's health, especially for people who have specific illnesses or healthcare needs. For example, there seems to be a link between the daily temperature and ED admissions for cardiovascular diseases or significant exacerbation of asthma in adults that visit ED [6] [7]. Generally speaking, regarding cardiovascular disorders, a worsening of the patient's well-being and cardiac arrests appear to be influenced by not only temperature but also stressors like humidity and atmospheric pressure [8] [9]. Moreover, there is also proof of links between air pollution and specific illnesses. Substances like $PM_{2.5}$, PM_{10}, NO_x, O_3 and SO_2 influence cardiac arrests [10], cardiac arrhythmia [11], cognitive decline in adult population [12], COVID-19 incidence [13], development of chronic kidney disease [14] or Type 2 diabetes [15]. $PM_{2.5}$ and PM_{10} are also linked to hospital admissions for cardiovascular [16] and respiratory diseases [17]. $PM_{2.5}$ levels also seem to be directly associated with increased daily ED visits for ulcerative colitis [18], while solar radiation is inversely associated with inflammatory bowel disease admissions [19]. There also seems to be a correlation between the number of hospitalised asthma patients and both weather (i.e., temperature and humidity) and pollution (i.e., $PM_{2.5}$, PM_{10} and NO_x) [20]. Finally, ML models (i.e., AutoRegressive Integrated Moving Average and Multilayer Perceptron) have also been applied to try to predict accesses to the ER by patients affected by infecting respiratory diseases after being exposed to $PM_{2.5}$ [21].

Some of these researches are based on long-term exposure to pollution (even 20-years long [12]), while others are based on a few days or even same day's exposure [13] [16] and some even on both [11]. Based on these literature pieces of evidence, trying to predict either all accesses to the ED or hospitalisation post-triage for specific illnesses, working on climate, pollution and historical accesses time-series belonging to the same area, seems feasible.

Each year, between 77000 and 80000 patients visit the ER of the biggest Brescia hospital [22] and 24% of them get admitted. This was the spark that ignited this work: trying to accurately predict future accesses to one of Brescia's EDs based on both historical and local meteorological and pollution

data.

This paper contains a description of the analysed materials and applied methods, i.e., the datasets and the ML approaches applied to them, in Section II, the reached results in Section III, a comment on them in Section IV and a few final remarks in Section V.

II. Materials and Methods

In this section, the study design, analysed datasets (both clinical and environmental data) and applied algorithms are described.

A. Study Design

This study primarily aims to provide a daily prediction of the amount of patients visiting the ER of a precise hospital in the city of Brescia, Italy. Also, a forecast of the number of hospitalised patients for specific disease classes has been attempted. This retrospective study was performed based on daily data (clinical and environmental) for a period from January 1, 2018, to December 31, 2022. A four-year (i.e., 2018–2021) dataset was used to train the forecasting models, while the remaining data were used to test its forecasting capability. The final dataset that is used to feed the predictive algorithms is a combination of the clinical and the environmental data.

B. Data Collection: Clinical Data

The original clinical dataset was given by a hospital in Brescia to GPI for research purposes. The dataset contained all anonymous ER access data for the period 2018-2022. For each access (i.e., a person on a specific day) there were as many rows as the exams the person had undergone; pre-processing was made in order to have only one row for each ED visit while maintaining the patient's data (like the date of ER visit, their age, sex and zip code of their home address, the list of medical exams they were subjected to and, in case they were hospitalised, their diagnosis as an ICD9-CM code).

The following is a description of this dataset.

TABLE I. Brief description of clinical data.

Year	*Total accesses*	*Median age*	*Male percentage*
2018	60176	55	49%
2019	60106	56	49%
2020	47205	58	52%
2021	49571	57	50%
2022	56631	56	51%

In 2018, 12% of patients were below 18 years old, 31% between 19 and 49, 23% between 50 and 69, 34% above 70.
In 2019, 11% of patients were below 18 years old, 30% between 19 and 49, 24% between 50 and 69, 35% above 70.
In 2020, 9% of patients were below 18 years old, 29% between 19 and 49, 27% between 50 and 69, 35% above 70.
In 2021, 10% of patients were below 18 years old, 30% between 19 and 49, 26% between 50 and 69, 34% above 70.
In 2022, 12% of patients were below 18 years old, 29% between 19 and 49, 25% between 50 and 69, 34% above 70.
Amongst the most recurrent diagnoses of the hospitalised patients, through all years, were pneumonia and chronic heart failure. Note that this dataset contains accesses of people living not only in Brescia but also in the province of Brescia and other places in Italy and abroad. However, what we included in our final dataset is:

- Daily number of accesses to the ER, limited to those patients coming only from the city of Brescia
- The rolling mean of the number of the same patients, applying a seven-day window for calculation.

The following is a description of the dataset restricted to Brescia.

TABLE II. Brief description of clinical data (city of Brescia).

Year	*Total accesses*	*Median age*	*Male percentage*
2018	10389	56	46%
2019	10963	58	47%
2020	9835	61	50%
2021	11082	60	49%
2022	12597	60	49%

In 2018, 11% of patients were below 18 years old, 30% between 19 and 49, 24% between 50 and 69, 35% above 70.
In 2019, 10% of patients were below 18 years old, 29% between 19 and 49, 24% between 50 and 69, 37% above 70.
In 2020, 8% of patients were below 18 years old, 27% between 19 and 49, 27% between 50 and 69, 38% above 70.
In 2021, 9% of patients were below 18 years old, 28% between 19 and 49, 25% between 50 and 69, 38% above 70.
In 2022, 11% of patients were below 18 years old, 27% between 19 and 49, 23% between 50 and 69, 39% above 70.

A little bit of contextualisation of this clinical dataset: it is important to note that the area around Brescia suffered in a substantial way from the outbreak of the COVID-19 pandemic and the number of cases affected by coronavirus pneumonia far exceeds the occurrences of any other diagnosis during 2020. It is possible to observe from these data, and this is something already reported in previous studies [23] [24], that the number of accesses to ER decreased significantly from 2019 to 2020: this is explainable because Italy was subjected to a strict lockdown for several months that year. Hence it was less likely, for example, for car accidents to happen or for people wearing masks to get the flu.

C. Data Collection: Environmental Data

The environmental data have been supplied by the startup Hypermeteo [25] under GPI's specific request to match the spatio-temporal dimension of the already-at-disposal clinical dataset. The environmental data are defined per day and zip code, guaranteeing spatial-temporal precision. These data are obtained employing a mathematical model with a resolution of $10km$x$10km$, corrected through normalisation and downscaling, applied to data by Lombardia's Regional Environmental Protection Agency (ARPA [26]) weather stations. While the

model was built for the entire Lombardia region, data were extracted for the province of Brescia only and, for this initial phase of the study, only data from the city of Brescia itself have been analysed.

The reported variables are:

- Temperature (min and max values) (T_{min}, T_{max} [$°C$])
- Humidity (min and max percentage values) (RH_{min}, RH_{max} [$\%$])
- Precipitations (Prec [mm])
- PM_{10} and $PM_{2.5}$ [$\mu g/m^3$]
- NO_x, SO_2 and O_3 [$\mu g/m^3$]
- Total solar irradiance (SSW_{tot}) [Wh/m^2].

For each variable, safety ranges, provided along with the dataset, were considered in order to give a label (i.e., zero or one) to each value, to indicate if a value could be considered safe or not. Depending on the type of variable, either lower or upper bounds were considered, as reported in Table I.

TABLE III. SAFETY RANGES FOR ENVIRONMENTAL VARIABLES.

Environmental variable	Lower and Upper Bounds	
	Min value	***Max value***
NO_x	25 $\mu g/m^3$	-
$PM_{2.5}$	15 $\mu g/m^3$	-
PM_{10}	45 $\mu g/m^3$	-
O_3	100 $\mu g/m^3$	-
SO_2	40 $\mu g/m^3$	-
T_{min}	-	-10 $°C$
T_{max}	35 $°C$	
RH_{min}	-	15 $\%$
RH_{max}	95 $\%$	-
Prec	-	10 mm
SSW_{tot}	-	8500 Wh/m^2

Regarding the dataset for the city of Brescia, the number of occurrences in which the data were out of range was computed. Occurrences are to be intended as a single day of the five years considered, per single zip code (Brescia has 15 different zip codes). In the 71% of occurrences, NO_x results out of range, it is the 60% of cases for $PM_{2.5}$, 20% for PM_{10}, 17% for the max humidity, 7.7% for the precipitations, 7.4% for O_3, 1.8% for the max temperature and 0 cases out of range for SO_2 and the min temperature.

The issue of having multiple rows of data for the same date (i.e., one row for each zip code) has been handled similarly as in a project [27] found during our bibliographic research: each environmental variable has been labelled with the zip code it is referred to, and it is used as a column with daily values, thus grouping all data belonging to the same date on one row. Again, a clarification on the context: the area surrounding Brescia is densely inhabited and industrialised, resulting in one of the most polluted areas in Europe [28].

D. Predictive Algorithm: Random Forest

In order to predict future ER accesses based on our clinical and environmental data, a Random Forest (RF) approach was implemented on Python applying the open-source library Scikit-learn [29]. This model was chosen based on an article [30] that applied it to a temperature prediction problem: the analogy with our dataset highlighted this approach as an interesting candidate for this type of analysis. RFs apply sequential splits to the data such that the separation is maximised in regards to a homogeneity criterion resulting in a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [3]. The random forest algorithm picks N random records from the dataset and builds a decision tree based on them, repeatedly for the chosen number of trees (in this case, 1000). The topic has been tackled as a regression problem as we have considered the target variable (i.e., daily accesses) as a continuous one.

Through the same library cited before, some metrics were computed to evaluate the results: the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE) and the Accuracy (Acc). Then, when the prediction of the number of daily hospital admissions for cardiovascular and respiratory pathologies was attempted, the Symmetric Mean Absolute Percentage Error (SMAPE) was computed. This analysis was applied expecting a more evident correlation between environmental, especially pollution, data. These pathological classes have been selected through their ICD9-CM codes.

The results that are reported in Subsection III-A, are based on different combinations of the datasets, as we applied the same model on the entirety of Brescia's data, only on the 2 most important features and only on cardiovascular and respiratory disorders data, respectively. In order to highlight a possible lag effect based on 1- and 5-day lag assumptions, which means that the observed data of previous days is used to predict the volume of patient access on a certain day, climate and pollution data were processed accordingly in order to create two analogous additional datasets.

The different analyses that were carried out, trying to improve the model's accuracy and potentially spot specific patterns, are divided into four cases:

- A; the RF algorithm was applied to the initial pre-processed dataset, then on 1-day and 5-day lagged data and, finally, only on the 2 most important features, as computed by the model
- B; analogous to A, but the rolling mean feature was discarded
- C; 1-day lagged data, no rolling mean, but the clinical data were reduced to only the part linked to hospitalised patients affected by cardiovascular pathologies, plus on the 2 most important features
- D; analogous to C, but the clinical data belonged to respiratory disorders.

Here, are reported the equations [33] for MAE (1), MAPE (2), SMAPE (3) and Acc (4):

$$MAE = \sum_{i=1}^{D} |x_i - y_i| \tag{1}$$

$$MAPE = \frac{100}{N} \sum_{i=0}^{N-1} \frac{y_i - \hat{y}_i}{y_i} \tag{2}$$

$$SMAPE = \frac{100}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(|A_t| + |F_t|)/2} \quad (3)$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

E. Predictive Algorithm: ARIMA

Trying to improve the results given by the algorithm described in Subsection II-D, a ML model for multivariate time-series prediction was applied to the same data. Specifically, an AutoRegressive Integrated Moving Average (ARIMA) model [31], a popular algorithm used in time series analysis and forecast, through the application of the auto-ARIMA process [32] in Python. The basic idea of the ARIMA model is to use a certain mathematical model to describe the random time series of the data, then predict the future values based on the past, and present values, through a so-called autoregression. An ARIMA (p, d, q) model can be described in the following equation (5).

$$(1 - \sum_{i=1}^{p}\varphi_i L^i)(1 - L)^d X_t = (1 + \sum_{i=1}^{q}\theta_i L^i)\varepsilon_t \quad (5)$$

where L represents the lag operator, p represents the number of autoregressive terms, q represents the number of moving average terms, d represents the degree of differencing and ϕ, θ and ϵ are relevant parameters.

The performance metrics applied to the model to evaluate its performances were MAPE (2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Akaike Information Criterion (AIC). Here, are reported the equations for MSE (6), RMSE (7) and AIC (8).

$$MSE = \sum_{i=1}^{D}(x_i - y_i)^2 \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2} \quad (7)$$

$$AIC = 2k - 2\ln(\hat{L}) \quad (8)$$

III. Results

In this section, the obtained preliminary results are reported. The algorithms have been fed with different datasets that always include only data related to patients whose home address' zip code is inside the city of Brescia.

A. Results: Random Forest

Following, a series of plots is reported: they represent the predicted values (plotted in violet) versus the actual values (plotted in blue) for the year 2022, coming from the different input datasets as explained in Subsection II-D.

First, the results of case A. Figure 1 displays the actual test values and the predicted ones for the 1-day lagged data.

Here, the obtained metrics for the 1-day lagged dataset (i.e., MAE and Acc) and for the 2 most important features, i.e., rolling mean and day (referring to the number of the day in a month), (i.e., MAE$_{mostimp}$ and $Acc\_{mostimp}$) are reported:

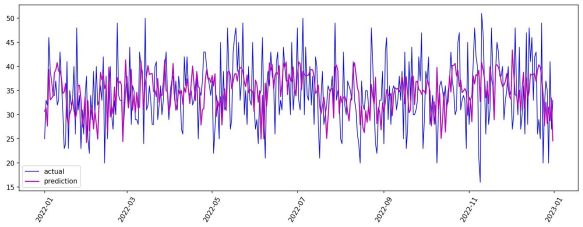


Figure 1. Random Forest's prediction and actual values for 1-day lagged data.

- MAE = 5.1
- Acc = 84.42%
- MAE$_{mostimp}$ = 5.57
- $Acc_{mostimp}$ = 82.63%

Now, the results of case B. Figure 2 displays the actual test values and the predicted ones for the 1-day lagged data missing the rolling mean.

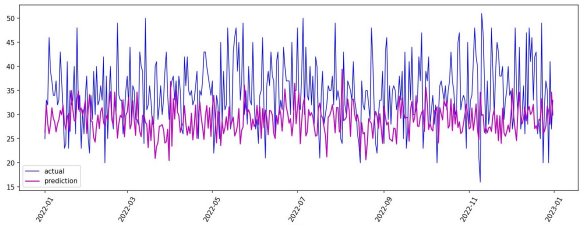


Figure 2. Random Forest's prediction and actual values without considering the rolling mean.

Here, the obtained metrics are reported (i.e., MAE$_{mostimp}$ and $Acc_{mostimp}$ refer to features day and month):

- MAE = 6.33
- Acc = 82.44%
- MAE$_{mostimp}$ = 7.49
- $Acc_{mostimp}$ = 79.29%

For case C, the actual and obtained predicted data are displayed in Figure 3.

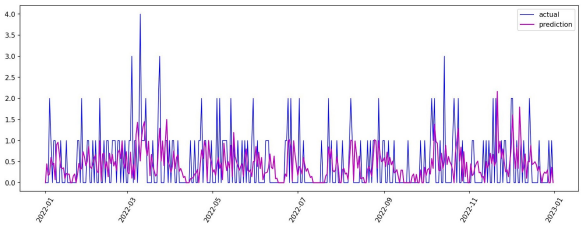


Figure 3. Random Forest's prediction and actual values for cardiovascular diseases' hospitalisations.

The obtained metrics were (with rolling mean and day as the most important features):

- MAE = 0.51
- MAE$_{mostimp}$ = 0.49

Case D's plot of predicted and actual values for respiratory pathologies is Figure 4.

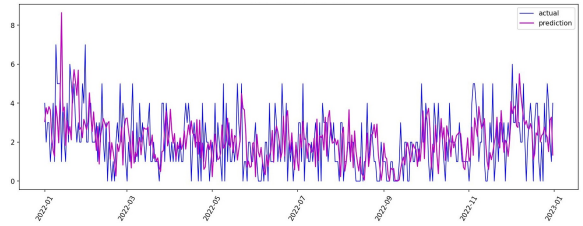


Figure 4. Random Forest's prediction and actual values for respiratory diseases' hospitalisations.

The obtained metrics were (with rolling mean and day as the most important features):

- MAE = 1.09
- SMAPE = 67.9%
- $\text{MAE}_{mostimp}$ = 1.23
- $\text{SMAPE}_{mostimp}$ = 74.9%

In this case, SMAPE was computed, instead of MAPE and Acc, due to the presence of 0 values in the test array.

B. Results: ARIMA

Here, the results obtained with the auto-ARIMA algorithm are shown: they represent the predicted values (plotted in violet) versus the actual values (plotted in green) for the year 2022. This prediction is obtained by feeding the initial dataset to the algorithm. The plot of the actual test values and the predicted ones for the same-day data is reported in Figure 5

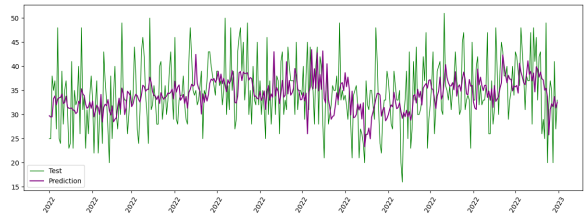


Figure 5. ARIMA's prediction and actual values for same-day data.

The obtained metrics were:

- MAPE = 15%
- MSE = 39.5
- RMSE = 6.3
- AIC = 9272.7

IV. Discussion

Results reported in Subsection III-A only refer to 1-day lagged data because, when the same process was applied to the same-day data and the 5-day lagged data, results were quite similar. Hence, in order to show the model performances, the former was chosen as it seemed to be the best logical approach. Results reported in Subsection III-B only refer to same-day data as it was the outcome of an early analysis of the ARIMA model application to these datasets, thus only the initial valuations have been implemented.

Visually comparing both models, predictions coming from the RF algorithm (Figure 1, 2, 3 and 4) appear to adhere better to the actual data when compared with the ARIMA one (Figure 5). However, even if the specific value is not always correctly predicted, the overall trend seems to be rightly predicted. This also happens while changing the considered features, despite removing historical data like the rolling mean (Figure 2), thus relying more on the environmental ones. In fact, the forecast values are underestimated, but the trend is followed quite well. Still, generally, the RF model also predicts peak values (Figure 4), i.e., surges in hospitalisations, quite aptly.

Please note that when analysing specific pathologies, the number of hospitalisations is limited to a few people every day and, sometimes, even none. This is particularly noticeable as, in this work, only the city of Brescia's data are used and it is more obvious for cardiovascular disorders rather than the respiratory ones, at least during the considered period.

Beyond the visual inspection, the metrics results reported in Subsection III-A show that the Acc for the RF model decreases when discarding the rolling mean as an input feature, but only of 1.98% and the error on the predicted number of accesses (i. e., MAE) increases from 5.1 to 6.33. This seems to suggest that when using the historical data through the rolling mean, the prediction could still be improved, but also that when this feature is ignored, the forecast performances do not dramatically worsen. Similar reasoning can be applied to the approach that uses the two features computed to be the most important ones, which behaves even less precisely.

Results for the RF application to cardiovascular and respiratory data seem to output better punctual predictions, but MAE values are smaller because of the lower values of daily hospitalisations (when compared to daily general accesses) and SMAPE is quite high. This could be due to the nature of the dataset itself as it is quite small. Regarding the ARIMA metrics reported in Subsection III-B, the listed AIC value is the one belonging to the best model identified by auto-ARIMA and the MAPE value represents a low, but acceptable accuracy. As expected by the visual inspection, though, MSE and RMSE values are not adequate.

Based on the aforementioned decrease in ER accesses during 2019 and 2020 due to the COVID-19 pandemic, an attempt at training the models only on 2018 and 2019 data (and still testing them on 2022 ones) was made, hoping to improve the preciseness of the predictions, but, surprisingly, in vain. In fact, the hypothesis was to discard the out-of-the-ordinary data so that the predictions computed merely on the historical data could be more precise. The results' worsening could be further evidence that the previously obtained results were not only due to historical data but also to environmental info which influences the correctness of the forecasting.

V. Conclusion and Future Work

This work represents a starting point towards the time-series analysis of historical and environmental data for the prediction of ER accesses and hospitalisations in a specific geographical area. The objective was only partially reached as this is a demanding field of application, but results were generally promising and, under these premises, a predictive analysis seems feasible. Considering that there are no other

truly comparable works in the international literature, these performances are even more encouraging. This being said, the obtained results cannot be generalised as they were achieved by analysing a period greatly made up of COVID-19-ridden years and a quite limited geographical area, so they can only be used to comment on this specific frame. The performances could dramatically differ if the analogous pre-processing and the same models were to be applied to other contexts or even just on a longer and more stable period.

Future developments of this work will, of course, include data belonging to the entire province of Brescia and a continuous search for more precise results, with the hope of moving to ever-growing datasets. It would also be interesting to test other ML algorithms or apply different pre-processing steps. Nevertheless, any attempt, whether it be successful or inconclusive, will still gather valuable insight on this yet to be delved into the topic and shed light on how our surrounding environment influences human health. This may be the offset of a new way of managing ER all over the world, monitoring entire populations and geographical areas, with the final objective of enabling a smart real-time predictive analysis able to improve the quality of healthcare and people's quality of life.

REFERENCES

[1] J. D. Sonis and B. A. White, "Optimizing patient experience in the emergency department", Emergency Medicine Clinics, vol. 38, no. 3, pp. 705–713, 2020.

[2] Z. Qiao et al., "Using machine learning approaches for emergency room visit prediction based on electronic health record data", Building continents of knowledge in Oceans of data: The future of co-created eHealth. IOS Press, pp. 111–115, 2018.

[3] Y. M. Chiu, J. Courteau, I. Dufour, A. Vanasse, and C. Hudon, "Machine learning to improve frequent emergency department use prediction: a retrospective cohort study", Scientific Reports, vol. 13, no. 1, p. 1981, 2023.

[4] A. Cameron, K. Rodgers, A. Ireland, R. Jamdar, and G. A. McKay, "A simple tool to predict admission at the time of triage", Emergency Medicine Journal, vol. 32, no. 3, pp. 174–179, 2015.

[5] R. Sánchez-Salmerón et al., "Machine learning methods applied to triage in emergency services: A systematic review", International Emergency Nursing, vol. 60, 2022.

[6] W. Zhu et al., "The effect and prediction of diurnal temperature range in high altitude area on outpatient and emergency room admissions for cardiovascular diseases", International Archives of Occupational and Environmental Health, vol. 94, no. 8, pp. 1783–1795, 2021.

[7] T. Abe et al., "The relationship of short-term air pollution and weather to ED visits for asthma in Japan", The American journal of emergency medicine, vol. 27, no. 2, pp. 153–159, 2009.

[8] D. Martinaitiene and N. Raskauskiene, "Weather-related subjective well-being in patients with coronary artery disease", International Journal of Biometeorology, vol. 65, pp. 1299–1312, 2021.

[9] M. Hensel et al., "Association between weather-related factors and cardiac arrest of presumed cardiac etiology: a prospective observational study based on out-of-hospital care data", Prehospital Emergency Care, vol. 22, no. 3, pp. 345–352, 2018.

[10] S. Kojima et al., "Fine particulate matter and out-of-hospital cardiac arrest of respiratory origin", European Respiratory Journal, vol. 57, no. 6, p. 2004299, 2021.

[11] M. A. Shahrbaf, M. A. Akbarzadeh, M. Tabary, and I. Khaheshi, "Air pollution and cardiac arrhythmias: a comprehensive review", Current Problems in Cardiology, vol. 46, no. 3, p. 100649, 2021.

[12] J. M. Delgado-Saborit et al., "A critical review of the epidemiological evidence of effects of air pollution on dementia, cognitive function and cognitive decline in adult population", Science of the Total Environment, vol. 757, p. 143734, 2021.

[13] S.-T. Zang et al., "Ambient air pollution and COVID-19 risk: evidence from 35 observational studies", Environmental research, vol. 204, p. 112065, 2022.

[14] M.-Y. Wu, W.-C. Lo, C.-T. Chao, M.-S. Wu, and C.-K. Chiang, "Association between air pollutants and development of chronic kidney disease: a systematic review and meta-analysis", Science of the Total Environment, vol. 706, p. 135522, 2020.

[15] Y. Li, L. Xu, Z. Shan, W. Teng, and C. Han, "Association between air pollution and type 2 diabetes: an updated review of the literature", Therapeutic Advances in Endocrinology and Metabolism, vol. 10, pp. 1-15, 2019.

[16] R. D. Brook et al., "Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association", Circulation, vol. 121, no. 21, pp. 2331–2378, 2010.

[17] F. Dominici et al., "Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases", Jama, vol. 295, no. 10, pp. 1127–1134, 2006.

[18] R. Duan et al., "Association between short-term exposure to fine particulate pollution and outpatient visits for ulcerative colitis in Beijing, China: A time–series study", Ecotoxicology and Environmental Safety, vol. 214, p. 112-116, 2021.

[19] F. Jaime et al., "Solar radiation is inversely associated with inflammatory bowel disease admissions", Scandinavian journal of gastroenterology, vol. 52, no. 6-7, pp. 730–737, 2017.

[20] C.-L. Chan et al., "A survey of ambulatory-treated asthma and correlation with weather and air pollution conditions within Taiwan during 2001–2010", Journal of asthma, vol. 56, no. 8, pp. 799–807, 2019.

[21] J. Lu et al., "Feasibility of machine learning methods for predicting hospital emergency room visits for respiratory diseases", Environmental Science and Pollution Research, vol. 28, pp. 29701–29709, 2021.

[22] https://civile.asst-spedalicivili.it/servizi/unitaoperative/unitaoperativefase02.aspx?ID =586 [Retrieved online: January 2024].

[23] F. Tartari, A. Guglielmo, F. Fuligni, and A. Pileri, "Changes in emergency service access after spread of COVID-19 across Italy", Journal of the European Academy of Dermatology and Venereology, vol. 34, no. 8, p. e350, 2020.

[24] T. Ferrari, C. Zengarini, F. Bardazzi, and A. Pileri, "In-depth, single-centre, analysis of changes in emergency service access after the spread of COVID-19 across Italy", Clinical and Experimental Dermatology, vol. 46, no. 8, pp. 1588–1589, 2021.

[25] https://www.hypermeteo.com/ [Retrieved online: January 2024].

[26] https://www.arpalombardia.it/ [Retrieved online: January 2024].

[27] https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather [Retrieved online: January 2024].

[28] S. Khomenko et al. "Premature mortality due to air pollution in European cities: a health impact assessment", The Lancet Planetary Health, vol. 5, no. 3, pp. e121–e134, 2021.

[29] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.ht [Retrieved online: January 2024].

[30] https://towardsdatascience.com/random-forest-in-python-24d0893d51c0 [Retrieved online: January 2024].

[31] https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html [Retrieved online: January 2024].

[32] https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html [Retrieved online: January 2024].

[33] G. Vishwakarma, A. Sonpal, and J. Hachmann, "Metrics for benchmarking and uncertainty quantification: Quality, applicability, and best practices for machine learning in chemistry", Trends in Chemistry, vol. 3, no. 2, pp. 155-156, 2021.

Medical Knowledge Harmonization: A Graph-based, Entity-Selective Approach to Multi-source Diagnoses

Andrea Bianchi
DISIM Department
University Of L'Aquila
L'Aquila, Italy
andrea.bianchi@graduate.univaq.it

Antinisca Di Marco
DISIM Department
University Of L'Aquila
L'Aquila, Italy
antinisca.dimarco@univaq.it

***Abstract*—The paper discusses a novel system for medical diagnostics that integrates patient data from various sources to address the fragmentation of healthcare information. By generating and merging knowledge graphs from raw medical texts focused on key biomedical entities (Gene, Disease, Chemical, Species, Mutation, Cell Type), the system facilitates a comprehensive understanding of a patient's medical history. It accurately extracts and connects critical entities, creating individual and combined knowledge graphs that elucidate a patient's medical journey. This approach helps bridge diagnostic gaps, offering a visual tool for practitioners to detect patterns and discrepancies in patient data. Despite limitations such as language dependency and validation scope, this system sets the stage for future enhancements toward a more universally accessible and clinically useful healthcare system.**

***Index Terms*—medical diagnostics, multi-source diagnosis**

I. Introduction

In modern healthcare systems, a patient often consults with multiple specialists across different institutions, leading to multiple diagnostic records. These records, though rich in information, can often be fragmented and inconsistent [1]. As a result, for chronic or complex illnesses, a single individual may have many diagnoses, sometimes different and spanning different time periods and institutions. While this multitude of data sources should, in theory, provide a comprehensive view of a patient's health, it often results in the opposite: a fragmented, and occasionally contradictory puzzle of information [2]. This overwhelming and fragmented landscape of patient data can lead to gaps in understanding, potentially causing misdiagnoses, redundant testing, and even treatment errors [3]. Knowledge graph (KG) is a systematic way to connect information and data points to knowledge. These graphs may effortlessly combine intricate patient data in the context of medical diagnostics, making them an appropriate solution for managing discussed challenges [4].

This paper introduces an approach to tackle the problem of multi-source diagnostic data integration, a process that involves combining diagnostic information from various healthcare sources to create a cohesive patient health profile. The problem is intriguing because resolving it has the potential to significantly enhance diagnostic accuracy and treatment efficacy. It's particularly vital in genetic information and rare diseases, where integrating scattered and specialized data can lead to breakthroughs in understanding and treatment. While previous efforts have made strides in improving data quality and developing data exchange standards, they often fall short in addressing the semantic integration of complex medical data comprehensively. Our work aims to bridge this gap by not only generating but also merging knowledge graphs from various diagnostic sources, thereby offering a panoramic and unified view of a patient's medical history. This approach stands to revolutionize how medical professionals access, interpret, and utilize patient data for more informed decision-making.

In light of these considerations, the primary contributions of this work are framed around three key research questions: (*RQI*) How can individual knowledge graphs be generated from raw medical texts? (*RQII*) What mechanism allows for the merging of these individual graphs while highlighting unique entities? And (*RQIII*) How can a visualization tool assist medical professionals in understanding a patient's comprehensive medical history? Addressing these questions, our paper outlines the methodology for generating and merging knowledge graphs, followed by an exploration of a visualization tool designed for medical professionals.

The paper is structured as follows: Section II reviews related works, while Section III presents the motivating scenario behind our work. The technical details of our approach are explored in Section IV while Section V and Section VI discuss the experimental settings and results, respectively. Section VII provides a discussion on our findings. Finally, we conclude and discuss future work in the Section VIII.

II. Related Works

Different approaches and goals have been seen in the field of building knowledge graphs from medical and biological texts. In order to provide a more comprehensive representation of medical situations, some research projects aim to augment textual data with multiple notations that include genetics, proteomics, symptoms, and more [5] [6]. Others are focused on

 ISBN: 978-1-68558-136-7

developing knowledge graphs that are specialised to particular illness types and provide in-depth insights into their complex dynamics [7]. Moreover, some initiatives, such as [8] and [9], aim to generate knowledge graphs straight from spoken dialogues or utterances recorded in-context clinical encounters. Authors of [8] proposed a method to construct a medical knowledge graph directly from clinical conversations between doctors and patients. Unlike this work, our approach aim at providing a unified visualization that emphasizes patient's whole medical journey rather than predictive analysis from singular clinical conversations. PrimeKG [5] serves as a multimodal knowledge graph for precision medicine, integrating data from 20 resources to offer insights across ten biological scales, from protein perturbations to therapeutic drug actions. [6] introduces the Clinical Knowledge Graph (CKG), an expansive platform designed to integrate diverse biomedical data, including proteomics, to facilitate precision medicine. CKG, encompassing over 16 million nodes and 220 million relationships, aims to represent experimental data, public databases, and literature while implementing advanced statistical and machine learning tools to enhance proteomics workflows. Differently from [5] and [6], our research is tailored towards unifying diagnostic data from multiple healthcare centres, providing a comprehensive visual picture of a patient's medical trajectory.

III. Motivating Scenario

Consider a scenario where distinct diagnostic reports, generated at different times and by different institutions, capture varied aspects of a patient's health. A report from one hospital might highlight specific findings that were either not observed or not considered pertinent in another [10], [11]. The proposed system ingests diagnostic texts from various sources and generates individual knowledge graphs. These graphs, each representing a unique diagnostic perspective, are then merged into a unified knowledge graph, as illustrated in Figure 1. This integrated visualization accentuates common entities and relationships using consistent colours and distinctly highlights unique entities or pieces of evidence from each diagnostic source. By offering this consolidated view, healthcare professionals are equipped with a panoramic understanding of an individual's health trajectory. This enables more informed decisions, ensures no detail is missed, and potentially avoids redundant or misguided medical interventions, ensuring the best possible patient care and improving personalized medicine [5].

IV. Medical Knowledge Harmonization

The goal of our system is to transform fragmented diagnostic texts into a unified knowledge graph, providing a holistic understanding of a patient's medical history. This transformation is achieved through a series of systematic steps, as depicted in Figure 1, General Workflow. The figure delineates our workflow through four pivotal macro-steps: 1) Named Entity Recognition (NER), where entities are identified from the raw texts; 2) Relationship Extraction (RE), where relationships between identified entities are extracted; 3) Single Source Graph Generation, which involves creating individual knowledge graphs for each diagnostic source of each patient; and 4) Knowledge Graph Integration, where these individual graphs are amalgamated into a unified, comprehensive knowledge graph.

A. Input Source Determination and Preprocessing

The system processes multiple diagnostic texts from varied healthcare environments (Figure 1, Tools), reflecting different stages of a patient's medical history. It operates in two modes: **Data Ingestion Mode**, which uses a structured dataset to generate and integrate knowledge graphs, and **Manual Mode**, where users manually input diagnostic reports for ad-hoc analysis. In Manual Mode, reports are uploaded to a specific folder (*diagnostic_reports*), and the system then extracts and integrates data into the knowledge graph, similar to the Data Ingestion Mode. This flexibility allows for both comprehensive and targeted analyses of patient diagnostics.

B. Entity Recognition and Normalization

Each diagnostic text (T1, T2, ... Tn) undergoes NER to identify medically relevant entities. This step utilizes NER techniques and tailored for medical and biological texts, ensuring accurate extraction of entities. For this critical task, our system employs BERN2 [12], a state-of-the-art tool in the biomedical domain, which is capable of recognizing and normalizing nine different entities: Gene, Disease, Chemical, Species, Mutation, Cell Line, Cell Type, DNA, and RNA. BERN2 adopts distinct strategies for multi-task NER, ensuring accurate extraction of entities by navigating through the intricate and domain-specific language of medical and biological texts. Subsequent to the entity recognition, BERN2 proceeds with the normalization of these entities, utilizing dedicated methods that enhance the precision and reliability of the identified entities within the diagnostic texts.

C. Relation Extraction

After the entities have been recognized and normalized, the system advances to the RE stage, which aims to decipher the relationships between the identified entities within the diagnostic text. For this endeavor, we use the capabilities of Bio_ClinicalBERT [13]. Bio_ClinicalBERT is a model developed for processing clinical text. It combines BioBERT's pretraining on biomedical literature with further training on MIMIC-III notes, a database of electronic health records from ICU patients. The model, trained on a variety of notes, is designed to capture the nuances of clinical language [14]. Despite not being originally designed to discern relationships between entities, the embeddings from Bio_ClinicalBERT, enriched with substantial biomedical and clinical contextual information, can be leveraged to infer potential relationships among the identified entities through a heuristic approach. It's worth noting that our experiments also leverage the MIMIC database, aligning our experimental setup with the intrinsic knowledge and understanding embedded within Bio_ClinicalBERT, thereby ensuring a coherent setting.

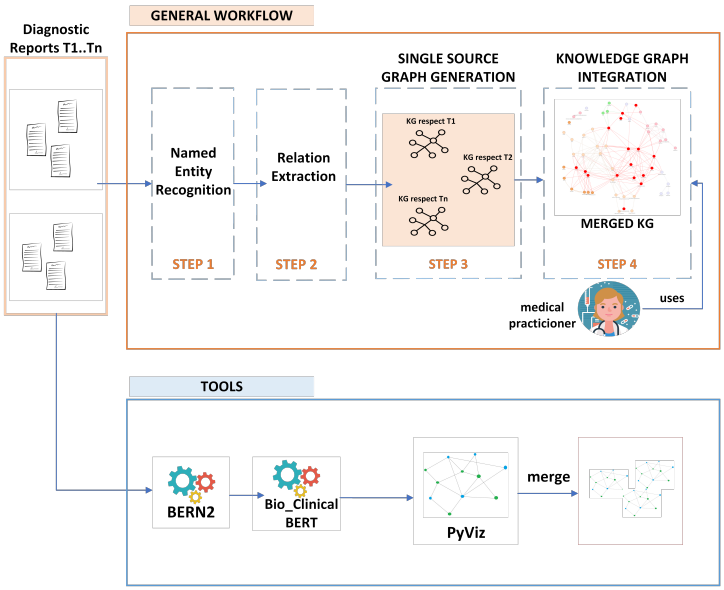


Figure. 1. High-Level Workflow for the system.

D. Knowledge Graph Generation

After extracting entities and their respective relationships, the system leverages on these to construct individual knowledge graphs for each diagnostic text, utilizing entities as nodes and their relationships as edges to graphically illustrate the information embedded within each text. Following the generation of these individual knowledge graphs, the system goes to the integration phase, wherein it amalgamates these multiple graphs into a unified knowledge graph. This consolidated graph stands as a coherent synthesis of information, amalgamating insights from all diagnostic sources and providing a comprehensive visual depiction of a patient's entire medical history. The visually integrated knowledge graph also highlights common entities and relationships with consistent colours.

V. Experimental Settings

Here we delve into the specifics of how our research was conducted, ensuring transparency and reproducibility.

A. Hardware Configuration

The study utilized the Caliban cluster at the University of L'Aquila, which has multiple nodes with 40 processing units for parallel execution in the "mpi" environment. The tests ran on a CentOS Linux 7.4.1708 system with an Intel Xeon E5-2698 v4 CPU at 2.20GHz and 141GB RAM.

B. Dataset

Our study uses the MIMIC-IV-Note dataset (version 2.2), featuring *331,794* discharge summaries and *2,321,355* radiology reports, all de-identified for patient confidentiality ([14]). We focus on discharge summaries to analyze patients' medical histories. For ethics and replication, our dataset and code are available at PhysioNet [14] and [15], respectively. To replicate our preprocessed dataset, specific steps are required.

- **Filtering for Discharge Notes:** We selected *discharge* notes from our database for their detailed summaries of hospital stays, including diagnoses, treatments, and medical histories.
- **Extracting *History of Present Illness*:** We used regex parsing to extract this section, providing a detailed narrative of the patient's condition at the time of a specific hospitalization.
- **Adapting for Multiple Hospitalizations:** For patients with several hospital stays, we adjusted the data structure to isolate each hospitalization, enabling analysis of medical condition progression across visits.
- **Selecting Patients with Multiple Diagnoses:** Our dataset only includes patients with multiple diagnoses to focus on complex or rare medical histories.

This process resulted in a dataset of 59,051 unique patients, each with detailed hospitalization records and *'History of Present Illness'*.

C. Software Configuration

For the **NER step** we used **(BERN2 [12])**, an advanced biomedical entity recognition service. The *load_bern2_model* function processes diagnostic reports to extract and structure named entities for further use. For the **RE step** we selected *(Bio_ClinicalBERT [13])*, a variant of BERT that's specialized for clinical and biological texts. This model's embeddings are pivotal in our approach to relation extraction. For each pair of entities in a report, embeddings are generated. Then, the cosine similarity between entity pairs determines if a relation exists, creating it if the similarity surpasses a predetermined threshold fixed to *0,85*. For the **Knowledge Graph Generation** we considered **PyVis** v.0.3.1 Accession date: 02/06/2023.

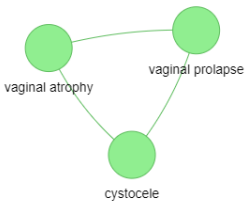


Figure. 2. Knowledge graph resulting from Medical Report 1 of the Experiment 1

The entities and relations derived from the aforementioned steps are organized into individual knowledge graphs using **Networkx** v.3.1 Accession date: 02/06/2023.

VI. Results

The experiment we show (detailed in Section VI-A) presents where the tool analyses the preprocessed dataset of 59,051 patients to produce knowledge graphs for numerous patients. This hypothetical situation might be similar to situations in which healthcare systems try to automatically create and preserve knowledge graphs for a large number of patients to aid in future consultations and plan creation.

A. Experiment 1: Automated Knowledge Graph Generation in Data Ingestion Mode

Example Case: Patient #10001876. Number of associated medical reports: 2.

Medical Report 1: Ms. ___ presented for evaluation of urinary complaints and after review of records and cystoscopy was diagnosed with a stage III cystocele and stage I vaginal prolapse, both of which were symptomatic. She also had severe vaginal atrophy despite being on Vagifem. Treatment options were reviewed for prolapse including no treatment, pessary, and surgery. She elected for surgical repair. All risks and benefits were reviewed with the patient and consent forms were signed.

Knowledge Graph for Medical Report 1. In Figure 2, we report the Individual Knowledge Graph generated by the system for Medical Report 1. Here only 3 interrelated entities have been extracted. Such graph is non informative with 7 entities and 13 relations.

Medical Report 2: She is a ___ patient who presents with ___ rectocele after having a sacral colpopexy and supracervical hysterectomy in ___ for uterine prolapse and cystocele. At that time, she had no rectocele at all. She has symptoms of bulge and pressure in the vagina that has gotten worse over the past few months. She also complains of feeling of incomplete emptying. She states that after she goes to the bathroom, she could go back and urinate some more. She had some frequency, urgency symptoms, which had resolved postoperatively. She also has resolved diarrhea after being started on Zenpep. She is followed by Dr. ___ and her fecal incontinence has resolved as well as resolved diarrhea."

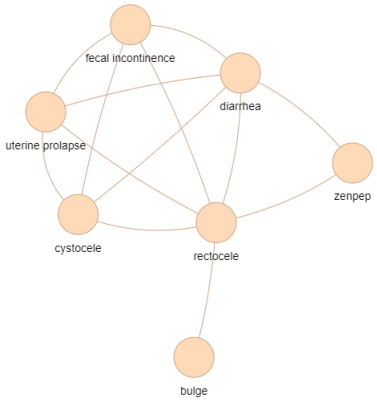


Figure. 3. Knowledge graph resulting from Medical Report 2 of Experiment 1

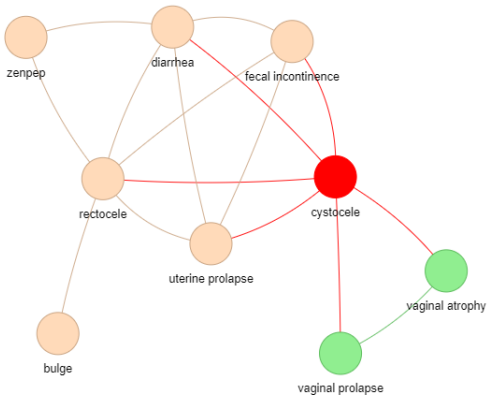


Figure. 4. Knowledge graph representing the merging of Medical Reports 1 and 2 for Experiment 1.

Knowledge Graph for Medical Report 2. In Figure 3 we report the Individual Knowledge Graph generated by the system for Medical Report 2

Merged Knowledge Graph of Experiment 1. Figure 4 shows the combined knowledge graph from Experiment 1, highlighting the shared entity *cystocele*, found in both reports, as a key connection point. This shared diagnosis suggests an ongoing or recurrent condition, emphasizing the importance of continuous monitoring and management. Recognizing such common conditions is essential for tracking disease progression or recurrence, aiding healthcare professionals in tailoring treatment plans to the patient's long-term medical history and current condition. Unique entities across reports, representing different medical conditions and treatments, are equally critical. For instance, *'vaginal prolapse'* noted in the first report, and *'rectocele'* and *'fecal incontinence'* in the second, highlight separate medical issues the patient has faced. These conditions—cystocele, vaginal prolapse, and fecal incontinence—are interconnected pelvic floor disorders. They involve the bladder bulging into the vagina, pelvic organ descent, and loss of bowel control, respectively, often due to weakened pelvic support ([16], [17], [18]). This information is vital for

TABLE I
COMPUTATIONAL TIMES.

Experiment	Time Required
Single patient (between 2-6 reports)	20-50 seconds
Entire dataset (59,051 patients)	12-14 days

TABLE II
SPACE USAGE.

Graph Type	Memory average	Memory - range	Memory (all patients)
Single Knowledge Graph	24 KB	[6 KB - 42 KB]	6.39 GB
Merged Knowledge Graph	56 KB	[8 KB - 110 KB]	3.9 GB

understanding the comprehensive scope of the patient's health challenges and planning appropriate interventions.

B. Computational Time and Space Usage

Creating knowledge graphs for each patient in our large dataset poses computational challenges. Our methodology accelerates graph generation, yet processing time escalates with dataset size, complexity of reports, and the quantity of entities and relationships. We utilized BERN2 API for NER, adhering to a 300 request limit per 100 seconds by incorporating 3-second pauses, prolonging processing for our dataset of 59,051 patients. The specific processing times are outlined in Table I. Storage requirements also significantly impact our experiments, with the space needed for individual and merged graphs dependent on the complexity and details of the diagnostic reports. Table II provides space usage statistics, showing that individual graphs require a total of 6.39 GB, while merged graphs need 3.9 GB.

VII. DISCUSSION

Addressing the complexity of healthcare information, our system autonomously creates and combines knowledge graphs from raw medical texts, navigating this crucial and challenging domain. Given the enormous variety of medical and biological entities present in healthcare, it was practical for us to narrow our primary attention to a small number of biomedical entities. This emphasis was seen in the studies, which showed the system's skill at locating, extracting, and connecting these chosen elements to create knowledge graphs that depict a clear and insightful narrative of a patient's medical journey. Focusing on a particular group of entities at this point allowed for deeper and more accurate knowledge as well as opened the door for methodical extension and inclusion of a wider variety of entities in the system's subsequent iterations. The experiments demonstrated the system's capability to accurately and coherently navigate medical texts, generating individual and merged knowledge graphs that highlight key entities and recurring illnesses, essential for understanding a patient's medical history and refining therapeutic strategies. The visualization tool emerged as a vital asset, offering medical professionals an intelligible visual narrative of a patient's medical journey, enhancing understanding and diagnostic ability.

A. Bridging Health Gaps: Societal Benefits of Comprehensive Medical Views

During brief appointments, some patients may find it difficult to remember and describe every medical exam, symptom, or medicine they have ever experienced (older people or people who are naturally reticent to retell every aspect of their medical history). Our system addresses these challenges by integrating multiple diagnostic reports into a unified visual representation. This ensures that every patient, irrespective of their background or communicative abilities, benefits from a comprehensive record that encapsulates their entire health journey.

B. Cost Efficiency

The proposed approach decreases the risk of unnecessary medical exams by giving a comprehensive perspective of a patient's health, which saves public and private money. Patients with complex medical histories, such as rare diseases, benefit most from the system since it makes sure they receive timely and effective care regardless of how many healthcare professionals they consult.

C. Global Scalability and Integration into Existing Infrastructure

The system showed excellent scalability, handling a dataset of 59,051 patients effectively, essential for managing the expanding volume of medical data. It's modular, allowing updates or replacements of components (e.g., entity extraction, relation prediction) without affecting the overall workflow.

D. Limitations and Threats to Validity

Input Accuracy: One of the foundational premises of our system is the reliance on accurate and relevant input. It's necessary that users (namely, doctors) provide diagnostic texts pertaining to the same patient. The system is designed to compare and integrate these texts, and any discrepancy in the input, such as including texts from unrelated patients, can lead to misleading results.

Natural Language Dependency: Our current implementation is tailored for the English language. This is largely because we utilize pre-trained tools, which are predominantly trained on English medical and biomedical terminologies. While the system demonstrates efficacy with English texts, its applicability could be limited in regions with different native languages. Expanding the system's capability to cater to diverse languages remains a future target.

Lack of Direct Baselines: It's challenging to compare our system directly with existing tools. While many tools extract entities from biomedical text, there are no tools aiming at integrating multiple texts into a unified knowledge graph.

E. Future Directions

As our system continues to evolve, one of our primary goals is to ensure its accessibility and usability worldwide. To achieve this, we are actively considering the incorporation of multilingual models, which would enable the system to

process and understand medical reports in various languages, catering to a global audience. Moreover, a promising frontier for our system lies in leveraging the intricate patterns within the knowledge graphs. Our vision is to utilize dedicated pattern recognition techniques that systematically analyze these graphs, pinpointing recurring sequences or clusters of entities and relations that could be indicative of specific medical conditions or trajectories [19]. For instance, by analyzing a vast number of knowledge graphs and tracing back the diagnostic journeys of patients with a particular condition, we might discern that certain entity relationships frequently precede the diagnosis of that condition [20].

VIII. Conclusion

Healthcare, at its core, revolves around accurate and timely information. In our study, we demonstrate the power of software engineering to bridge gaps, connect dots, and provide a comprehensive view of a patient's medical journey. By integrating fragmented medical reports into a unified knowledge graph, we ensure that no detail is missed. This holistic approach amplifies the quality of care, particularly for those who might struggle to articulate their medical experiences. This research underscores the synergy between software engineering and medical informatics, demonstrated through a system adept at autonomously generating and merging knowledge graphs from medical texts. The targeted focus on specific biomedical entities showcased the system's precision in narrating a patient's medical journey. The experiments reflected not only the accuracy and utility of this system but also its potential to significantly impact healthcare by aiding in timely and informed decision-making. The potential healthcare ramifications are profound. By reducing redundant medical exams, we envision a path towards more efficient and cost-effective healthcare. Moreover, we are committed to utilizing the knowledge graphs to gain valuable insights, which will help us develop proactive healthcare strategies and enable early interventions.

Acknowledgments

European Union - NextGenerationEU - National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) - Project: "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013 - Avviso n. 3264 del 28/12/2021. LIFEMAP-Dalla patologia pediatrica alle malattie cardiovascolari e neoplastiche nell'adulto: mappatura genomica per la medicina e prevenzione personalizzata Traiettoria 3 "Medicina rigenerativa, predittiva e personalizzata" - Linea di azione 3.1 "Creazione di un programma di medicina di precisione per la mappatura del genoma umano su scala nazionale" of the Ministry of Health.

References

[1] F. C. Bourgeois, K. L. Olson, and K. D. Mandl, "Patients treated at multiple acute health care facilities: quantifying information fragmentation," *Archives of internal medicine*, vol. 170, no. 22, pp. 1989–1995, 2010.

[2] W.-Q. Wei, C. L. Leibson, J. E. Ransom, A. N. Kho, P. J. Caraballo, H. S. Chai, B. P. Yawn, J. A. Pacheco, and C. G. Chute, "Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 219–224, 2012.

[3] D. Dong, R. Y.-N. Chung, R. H. Chan, S. Gong, and R. H. Xu, "Why is misdiagnosis more likely among some people with rare diseases than others? insights from a population-based cross-sectional study in china," *Orphanet journal of rare diseases*, vol. 15, no. 1, pp. 1–12, 2020.

[4] L. Murali, G. Gopakumar, D. M. Viswanathan, and P. Nedungadi, "Towards electronic health record-based medical knowledge graph construction, completion, and applications: A literature study," *Journal of Biomedical Informatics*, p. 104403, 2023.

[5] P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Scientific Data*, vol. 10, no. 1, p. 67, 2023.

[6] A. Santos, A. R. Colaço, A. B. Nielsen, L. Niu, P. E. Geyer, F. Coscia, N. J. W. Albrechtsen, F. Mundt, L. J. Jensen, and M. Mann, "Clinical knowledge graph integrates proteomics data into clinical decision-making," *bioRxiv*, pp. 2020–05, 2020.

[7] A. Rossanez, J. C. Dos Reis, R. d. S. Torres, and H. de Ribaupierre, "Kgen: a knowledge graph generator from biomedical scientific literature," *BMC medical informatics and decision making*, vol. 20, no. 4, pp. 1–24, 2020.

[8] R. Kulkarni and Y. Haribhakta, "Building the knowledge graph from medical conversational text data and its applications," in *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. IEEE, 2022, pp. 1508–1513.

[9] M. R. Kamdar, W. Dowling, M. Carroll, C. Fitzgerald, S. Pal, S. Ross, K. Scranton, D. Henke, and M. Samarasinghe, "A healthcare knowledge graph-based approach to enable focused clinical search." in *ISWC (Posters/Demos/Industry)*, 2021.

[10] M. Schmidli, "Outcome of patients with acute coronary syndrome in hospitals of different sizes. a report from the amis plus registry," *Swiss medical weekly*, vol. 140, no. 2122, pp. 314–322, 2010.

[11] T. Hewitt, S. Chreim, and A. Forster, "Incident reporting systems: a comparative study of two hospital divisions," *Archives of Public Health*, vol. 74, no. 1, pp. 1–19, 2016.

[12] D. Kim, J. Lee, C. H. So, H. Jeon, M. Jeong, Y. Choi, W. Yoon, M. Sung, and J. Kang, "A neural named entity recognition and multi-type normalization tool for biomedical text mining," *IEEE Access*, vol. 7, pp. 73 729–73 740, 2019.

[13] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[14] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific data*, vol. 10, no. 1, p. 1, 2023.

[15] B. Andrea, "Medical Knowledge Harmonization: A Graph-based, Entity-Selective Approach to Multi-source Diagnoses," https://github.com/anbianchi/knowledge_frombio, [Accessed 27-02-2024].

[16] C. Aboseif and P. Liu, "Pelvic organ prolapse," 2020.

[17] C. Reisenauer, A. Kirschniak, U. Drews, and D. Wallwiener, "Anatomical conditions for pelvic floor reconstruction with polypropylene implant and its application for the treatment of vaginal prolapse," *European Journal of Obstetrics & Gynecology and reproductive Biology*, vol. 131, no. 2, pp. 214–225, 2007.

[18] P. Abrams, K.-E. Andersson, A. Apostolidis, L. Birder, D. Bliss, L. Brubaker, L. Cardozo, D. Castro-Diaz, P. O'connell, A. Cottenden *et al.*, "6th international consultation on incontinence. recommendations of the international scientific committee: evaluation and treatment of urinary incontinence, pelvic organ prolapse and faecal incontinence," *Neurourology and urodynamics*, vol. 37, no. 7, pp. 2271–2272, 2018.

[19] X. Tao, T. Pham, J. Zhang, J. Yong, W. P. Goh, W. Zhang, and Y. Cai, "Mining health knowledge graph for health risk prediction," *World Wide Web*, vol. 23, pp. 2341–2362, 2020.

[20] H. Wang, X. Miao, and P. Yang, "Design and implementation of personal health record systems based on knowledge graph," in *2018 9th international conference on information technology in medicine and education (ITME)*. IEEE, 2018, pp. 133–136.

Evaluating Text Pre-Processing Strategies for Clinical Document Classification with BERT

1st Sarah Miller, 2nd Serge Sharoff, 2nd Geoffrey Hall
UKRI CDT in AI for Medical Diagnosis and Care
School of Computing, University of Leeds
Leeds, UK
Email:scslmi@leeds.ac.uk, s.sharoff@leeds.ac.uk, g.hall@leeds.ac.uk

3rd Prabhu Arumugam
Multi-Modal Team
Genomics England
London, UK
Email:Prabhu.Arumugam@genomicsengland.co.uk

***Abstract*—In many Natural Language Processing (NLP) tasks, Bidirectional Encoder Representations from Transformers BERT and BERT-based techniques have produced state of the art results. However, this increase in performance comes with a caveat, limitations in the size of the text input the model can process. There are few studies that discuss the constraints of BERTs input length in the context of clinical documents, and as a result, little is known about how effective BERT is in this regard. To overcome these constraints, we investigate techniques for modifying the input text size of pathology report documents. By utilizing various BERT variants, we evaluate these approaches and examine the relative significance of domain specificity versus generic vocabulary training. We demonstrate that BERT models trained on domain knowledge outperform the vocabulary of standard models. In the process of classifying a set of variable-length pathology report texts, BERTs standard truncation approach, which removes text longer than the maximum, performs as well as more sophisticated text pre-processing techniques.**

***Index Terms*—BERT; Clinical Text; Natural Language Processing; Text classification.**

I. Introduction

An essential task that supports clinical workflows throughout health services is information extraction from clinical text documents. Healthcare providers currently invest considerable time and money for clinical specialists to complete this labor-intensive manual task. Automating this process with Natural Language Processing (NLP) has the potential to deliver efficiencies, saving both time and money [1].

Bidirectional Encoder Representations from Transformers (BERT) and BERT-based techniques have shown to deliver notable results across many NLP tasks [2][3]. However, adopting BERT base methods for use with clinical documents presents challenges (1) there is a limit to the input text size the model can process, and (2) they can be computationally demanding, especially during training. BERTs impressive performance can be attributed to its attention mechanism [3][4]. However, what makes BERT so powerful also contributes to its weakness. BERTs attention mechanism scales quadratically and thus limits the size of text input that can be processed by even the most advance computer hardware [5].

Unfortunately, clinical text documents often exceed BERTs maximum input. The maximum input size a BERT model can process is 512 tokens. Tokens are word representations BERT accepts as input, and tokens are not equivalent to words. During the tokenization process words can be split into multiple tokens, therefore, the word count of a document can not be used to determine input size. To address the differences in word to token ratio the inputs into BERT need to be pre-processed or the model architecture needs to be changed to accommodate longer sequences. In this paper, we look to assess the former. Clinical documents are not constrained to a structured format and information included in the texts is at the behest of whomever completes it. Some clinicians are very concise giving all key information in short sentences, whereas some will provide a lengthier description, each approach is clinically valid, but it does present challenges when pre-processing clinical texts [1]. Pre-processing clinical documents with varying formats when there is a limitation on how much of the text can be used is one of those challenges. Key information is distributed throughout documents at varied intervals, and when pre-processing the texts into sections it is difficult to know which sections of the text best contains the text required for classifications.

In our experiments, we evaluate four different text pre-processing strategies to investigate these challenges. We use three variants of BERT models on a multi-label clinical document classification task, using a set of cancer pathology reports from the Genomics England research environment [6]. To the best of our knowledge, there is only one study that investigates the impact of BERTs input size using pathology reports and our study advances their techniques. Our study is also the only study in this area that offers insight into how varying text sequence sizes influences results. The remainder of this article is structured as follows. Section II describes related work on BERT models and the input size limitations for clinical documents. Section III provides an explanation of the dataset and methods used in this study. In Section IV, we present the results of our experiments, and we conclude our findings in Section V.

II. Related Work

Research addressing the input size limitations of BERT has not received much attention in the clinical domain. The automation of ICD coding is the common goal of the few studies in this field and except for one study, all use the MIMIC-III database [7] discharge summaries for their tasks. However,

 ISBN: 978-1-68558-136-7

the results produced across these studies are not entirely consistent. For instance, even after using text pre-processing techniques to overcome BERTs input size constraints in [8][9] the authors discover that simpler networks perform better than BERT. Contrastingly, in [10][11] the authors find that BERT outperforms the simpler models when modifying for input size.

The text pre-processing methods used in these studies follow two approaches (1) truncation (from the right) of any text that exceeds the maximum input size or (2) hierarchical text pre-processing which involves splitting the text into n length segments with or without overlapping. The model individually processes each of the document segments, and to get the classification results for a document in its entirety, each of the segment outputs are combined using either a pooling or attention-based method.

To the best of our knowledge, only one study has investigated how input size restrictions affects other kinds of clinical texts. In [8] the authors use BlueBERT [12] to classify a set of cancer pathology reports as well as the MIMIC-III discharge summaries. They do not use the pathology reports for the ICD coding task. The pathology reports have a set of six document labels, but rather than using a multi-label classification approach, they train six individual models, one for each of the labels. Unlike the results produced for the ICD coding of MIMIC-III discharge summaries, there is no significant difference between BlueBERT, a CNN, and a HiSAN network when classifying the pathology reports. However, the authors in [8] only assess the models trained on the pathology reports using the hierarchical text pre-processing method and a single variant of BERT, BlueBERT, in their experiments.

Outside of the clinical domain there is one in-depth study that explores strategies to adapt BERT for long document classification. In [13] the authors use the standard BERT model to classify several non-clinical datasets and they find that taking the first 128 tokens and last 382 tokens of each document produces the best overall results. In [8] the authors argue this approach may not translate well to the clinical domain but they do not assess this method in any of their experiments. Therefore, in this paper we aim to fill in the gaps between these studies, by systematically investigating how to adapt BERT for the classification of pathology report texts irrespective of their length, and how different variants of BERT perform with the adaptations.

III. Methodology

In this section we present the techniques used in this study. First, we describe the dataset, secondly the models hyperparameters and tokenization settings, and lastly the text pre-processing strategies used for managing longer texts.

A. Dataset

The dataset used in the experiments is a curated dataset taken from the Genomics England research environment. In the dataset there are 15,825 plain text pathology reports for 5413 participants registered on the 100k genome project. The dataset contains reports for participants with three common types of cancer: breast, colorectal, and lung. Classification labels are provided by linking associated clinical records with the date and a tumour id. The dataset is multi-label and multi-class containing a total of 13 classes. The classes in the dataset were transformed into a multi-label set of features to make model training more efficient. Table I displays the dataset features and the distribution. The data is split into a training set of 7753, a validation set of 4748, and a test set of 3324.

B. Models and Hyperprameters

We installed three BERT models from the Huggingface model hub and followed the transfer learning approach. For sequence classification tasks in a multi-label setting, we use a sequence classification instance of the BERT models initialized with pre-trained parameters and fine-tune them for our task. For information on fine-tuning BERT models, we refer readers to resources available in [3][14]. The models used in this study are: (1) BERT-base-uncased [3] implemented as a baseline to compare the performance of the generic BERT vocabulary to clinical ones. (2) Bio_ClinicalBERT [15] which we opted to use because it has been pre-trained using all of PubMed and all MIMIC-III texts, rather than BlueBERT that has been trained with less of the data in both these datasets. (3) BiomedBERT (abstracts + full text) [16] is a model that is pretrained on just PubMed. However, the authors claim it is still superior at biomedical NLP tasks because of its succinct medical vocabulary for tokenization. To perform the document classifications

TABLE I
DATASET FEATURE DISTRIBUTION

Column	**Dataset distribution per label/class label**		
Label	***Features***	***Reports Per Class***	***Total Reports***
Disease Type	Breast	7767	15825
	Colorectal	6389	
	Lung	1668	
Histology Code	80703	985	15825
	81403	6664	
	84803	628	
	85003	6310	
	84803	1238	
Grade	80703	985	15825
	81403	6664	
	84803	628	
	85003	6310	
	84803	1238	

the pathology reports are fed into each of the BERT models, and it is the hidden state h, of the special [CLS] token, produced by BERT, which provides the classification. Because the dataset is multi-label the h[CLS] token, the models document representation, is passed through a sigmoid activation function to produce probabilities for each of the class labels. For further information regarding the [CLS] and other special tokens we refer readers to [3][17]. All three models are trained using an AWS Sagemaker ml.p3.2xlarge instance. Throughout literature training parameters vary for BERT models but for our experiments we opted for 3 epochs, because when we increased this value there was minimal to no difference gained in performance. Likewise for selecting the batch size and learning rate, we found that a batch size of 16 and a learning rate of 3-5e, using an Adam optimizer were the most optimal settings for our task.

C. BERT Tokenization

The BERT tokenizer converts text sequences into word piece tokens. Word piece tokens are words that have been split into segments. For example, the words learning and learned become learn #ing and learn #ed, making each of these words worth 2 tokens, or 4 tokens in total. The word to token ratio given throughout literature is approx., 400 words = 512 tokens and because the word to token limit can only be approximated, we split documents using the token length.

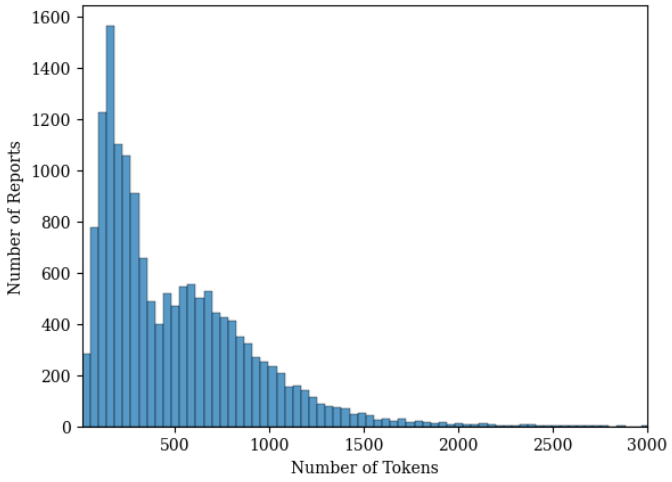


Fig. 1. Distribution of Report Token Lengths

To achieve this the pathology reports are passed through the BERT tokenizer to split the words in each report into their tokenized form. Fig 1 shows the distribution of token lengths for the pathology reports in the dataset. The reports vary in length with the shortest being just 10 tokens and the longest 5372. The mean token length for the dataset is 501, and at least 25% of the reports exceed 700 tokens. If a report is under the maximum it is processed in full. Wherever a report exceeds the maximum it is the count of the tokens that are used to split documents in the text pre-processing strategies.

D. Text Pre-Processing Strategies

a) Right and Left Truncation: An approach for handling sequences longer than 512 tokens is to implement a truncation strategy. BERT tokenizers take parameters for the sequence length and the position for truncation. BERT tokenizers offer either left or right truncation. The default setting is from the right and any tokens exceeding the specified length will be cut off from the right-hand side of the sequence. Likewise, with left truncation anything over the maximum is removed, but in this instance, it is removed from the left, from the beginning rather than the end of the sequence. In our experiments we adopt both approaches to truncation, and we use the maximum sequence length of 512.

b) Left+Right Truncate the Middle: Key information is said to be located at the beginning and end of a document. To investigate this further we follow the approach taken by the authors in [13] and take token segments from the beginning and end of the document, and concatenate them. For any document that exceeds the maximum sequence length of 512 we take the first 128 tokens of the document and the last 382, taking 510 tokens in total, leaving room for BERT special tokens. Any text/tokens in the document that fall in between these values are removed.

c) Hierarchical Text Pre-processing: Hierarchical text pre-processing is where long documents are broken up into segments. In this study any pathology report document exceeding the maximum input length is segmented into n=length/510 tokens. Each segment is prefixed with a [CLS] token and appended with a [SEP] token so they are 512 in length. Each segment is processed by the model following the fine-tuning approach. At the output stage each individual segment has a h[CLS] representation, and we apply mean pooling to combine the h[CLS] representations of all the segments giving a single output and the mean of the probabilities for the whole document.

E. Evaluation Metrics

The most commonly applied metrics in literature for evaluating NLP classification models are Accuracy, F1, and ROC-AUC scores [18]. For example, a popular set of NLP tasks for bench-marking models is GLUE [19] where the majority of tasks use Accuracy and or F1 for evaluation [18]. In the studies we reviewed F1, and ROC-AUC are the metrics reported. There is debate amongst the studies we reviewed which F1 metric is the most relevant, some favor macro F1, and other micro F1 scores for multi-label scenarios. In our experiments, we report micro F1, macro F1, and ROC-AUC in line with current literature for comparison.

IV. Experimental Results

The results in Table II and Table III are used to address three key questions: (1) how well the baseline model with a standard vocabulary compares to the domain trained models. (2) are there differences in performance between the two clinically trained models, and (3) how does text pre-processing to manage input sequence length impact classification performance.

TABLE II
DOCUMENT CLASSIFICATION RESULTS

Model	Model Classification Results			
	Text Strategy	*MicroF1*	*MacroF1*	*ROC-AUC*
BERT-base	Right Truncation	0.84	0.59	0.89
BERT-base	Left Truncation	0.82	0.52	0.88
BERT-base	Left+Right	**0.84**	**0.64**	**0.90**
BERT-base	H Mean Pooling	0.84	0.61	0.89
Bio_CBERT	Right Truncation	0.82	0.52	0.88
Bio_CBERT	Left Truncation	**0.84**	**0.67**	**0.89**
Bio_CBERT	Left+Right	0.84	0.62	0.89
Bio_CBERT	H Mean Pooling	0.84	0.63	0.89
BioMBERT	Right Truncation	0.88	0.69	0.92
BioMBERT	Left Truncation	**0.89**	**0.74**	**0.93**
BioMBERT	Left+Right	0.86	0.67	0.90
BioMBERT	H Mean Pooling	**0.90**	**0.74**	**0.93**

Model names abbreviated e.g., Bio_CBERT = Bio_ClinicalBERT

In respect to the first question, the clinical models do have an increase in performance compared to the BERT-base-uncased model. Confirming that domain specific models can offer an increase in performance when performing clinical NLP tasks. To answer question two there is a difference in performance between Bio_ClinicalBERT and BiomedBERT. This supports the studies claims that the BiomedBERT vocabulary is superior to other clinical variants even when they have been trained with more data. The BiomedBERT tokenizer is said to produce fewer word piece tokens than the other models and they attribute this to why it performs better, suggesting quality over quantity of data for the models training.

To address the final question, the BERT-base-uncased model has a slight increase in performance when using the Left+Right text pre-processing strategy. This reflects the results found by the authors in [13], but it is not reproduced in the results from the clinical models. The clinical models show minor differences but offer a slight increase in performance when using the left truncation and hierarchical mean pooling strategies (referred to as H Mean Pooling in Table II). Some pathology reports contain a summary of the key points of the investigation at the end of the report. Both favored text pre-processing strategies for the clinical models include the end of the document and could attribute to the increase in performance when using those strategies. To further address question three we investigate how the truncation of text has affected results by looking at the results over different document length distributions. Table III shows the results of the classifications across different subsets of document length. In Table III we have split the documents into groups using their original token lengths, prior to truncation, e.g., $\geq$1000 = documents with more than 1k tokens, and $\geq$512 $\leq$1000 are documents that have a token length greater than 512 but less than 1000 etc. We then group them also by the text pre-processing strategy used. What the results in Table III demonstrate is that there is a drop in performance with documents exceeding 1000 tokens. This is as expected, because these documents are subject to the most data loss, +50% of the data in these documents is removed. Longer documents contain key information throughout the length of the text, it is unlikely that it is all contained within the selected section, resulting in lost information required by the classifier. The results in Table III also reveal that there is a drop in performance that occurs for documents with token counts under the maximum limit. When the token counts drop below 250, these much shorter documents contain less information. They are lacking the data required for the successful classification of all the document labels. Thus, the shorter documents are also subject to data loss but in this instance because the clinician has perhaps missed information by being too concise. Changes in performance for texts with the highest and lowest token counts are observed across each of the text pre-processing strategies with BiomedBERT and truncation from the left providing the highest overall scores.

V. Conclusion and Future Work

In this study we have investigated how BERTs limitations in input size influences the classification of plain text pathology report documents. We find that there are performance increases when using a domain specific model for the task, and that not all domain model vocabularies are created equal. Similarly, to the other studies we reviewed the hierarchical text pre-processing approach does offer slightly better performance than the standard truncates from the right method. However, we also observed that for the pathology reports taking just the end of the text, truncation from the left performs just as well, and it is also a much faster method. Whilst our results are not entirely comparable to the results in [8], our models achieved higher macro F1 scores when classifying the pathology reports. Something that this study has highlighted is that the input length of a document is not just a factor when it is significantly longer than the maximum, but also when it is much shorter, and information is thus potentially missing. Pathology reports and other similar clinical texts are variable by nature. There are many factors at play that will dictate the content and length of clinical texts and because there is no current unified format or structure there is no guarantee that all information is recorded adequately. To address variations in the format of pathology reports, adopting a standardised approach could improve data quality for both clinicians and subsequent analyses. However, overall, the BERT models in this study performed well irrespective of the variations.

As previously addressed, currently there are limited studies for clinical document classification with BERT models. The ones that do exist use a limited set of documents from the

TABLE III
MACRO-F1 SCORES FOR CLASSIFICATIONS BY TOKEN LENGTH DISTRIBUTION

Text Pre-processing Strategy	**Macro F1 Scores for Token Length Evaluation**		
+ Token Length Distribution	***BERT-base***	***Bio_ClinicalBERT***	***BiomedBERT***
Right≥1000	0.57	0.52	0.66
Right ≥512 ≤1000	0.60	0.53	0.72
Right ≤512 ≥250	0.60	0.52	0.70
Right ≤250	0.57	0.51	0.68
Left ≥1000	0.51	0.60	0.72
Left ≥512 ≤1000	0.53	0.67	**0.76**
Left ≤12 ≥250	0.52	0.69	**0.77**
Left ≤250	0.51	0.66	0.72
Left+Right≥1000	0.58	0.60	0.62
Left+Right ≥512 ≤1000	0.65	0.63	0.70
Left+Right ≤512 ≥250	0.65	0.62	0.68
Left+Right ≤250	0.62	0.61	0.65

MIMIC-III database, and as discussed by [1] this does not provide a comparable enough view of this task. There needs to be more research using a variety of sources and use cases before the limitations of BERT models for clinical document classification can fully be established.

Future work will look at multi-task learning with BERT models and expanding the feature set of the dataset used in this study. Only a subset of the document features available for classification in the Genomics England research environment was used for this study, and there are potential further analyses, with a wider set of feature labels. BERT models are Deep Learning model architectures that are somewhat of a black box [20] and investigating the models output using explainability methods is also a future direction this research could take.

ACKNOWLEDGMENT

The research in this paper is part of a PhD project funded by the UKRI Center for Doctoral Training in Artificial Intelligence for Medical Diagnosis and Care (Project Reference: EP/S024336/1). Secondly, this work has also been supported by the Multi-modal team at Genomics England and we would like to give thanks to all team members for their invaluable knowledge and support, and lastly the research was made possible through access to data and findings in the National Genomic Research Library via the Genomics England Research Environment.

REFERENCES

[1] H. Dong et al., “Automated clinical coding: what, why, and where we are?”, npj Digit. Med., vol. 5, no. 1, pp. 1–8, 2022, doi: 10.1038/s41746-022-00705-7.

[2] I. Chalkidis et al., “An Empirical Study on Large-Scale Multi-Label Text Classification Including Few and Zero-Shot Labels,” CoRR, vol. 2010.01653, pp. 7503–7515, 2020, doi: 10.18653/v1/2020.emnlp-main.607.

[3] J. Devlin, M-W. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, vol. 1, no. Mlm, pp. 4171–4186. doi: https://doi.org/10.18653/v1/N19-1423.

[4] A. Vaswani et al., “Attention is all you need”, Proc. 31st Int. Conf. Neural Inf. Process. Syst., vol. 2017-Decem, no. Nips, pp. 6000–6010, 2017, doi: 10.5555/3295222.3295349.

[5] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The Long-Document Transformer”, ArXiv, vol. 2004.05150, 2020.

[6] M. Caulfield et al., “National Genomic Research Library.” Genomics England, London, UK, 2020. doi: 10.6084/m9.figshare.4530893.v7.

[7] A. E. W. Johnson et al., “MIMIC-III, a freely accessible critical care database,” Sci Data, vol. 3, p. 160035, 2016, doi: 10.1038/sdata.2016.35.

[8] S. Gao et al., “Limitations of Transformers on Clinical Text Classification,” IEEE J. Biomed. Heal. Informatics, vol. 25, no. 9, pp. 3596–3607, 2021, doi: 10.1109/JBHI.2021.3062322.

[9] [1] S. Ji, M. Hölttä, and P. Marttinen, “Does the magic of

BERT apply to medical code assignment? A quantitative study", Comput. Biol. Med., vol. 139, pp. 1–13, 2021, doi: 10.1016/j.compbiomed.2021.104998.

[10] C. W. Huang, S. C. Tsai, and Y. N. Chen, "PLM-ICD: Automatic ICD Coding with Pretrained Language Models", Clin. 2022 - 4th Work. Clin. Nat. Lang. Process. Proc., pp. 10–20, 2022, doi: 10.18653/v1/2022.clinicalnlp-1.2.

[11] A. Afkanpour et al., "BERT for Long Documents: A Case Study of Automated ICD Coding," LOUHI 2022 - 13th Int. Work. Heal. Text Min. Inf. Anal. Proc. Work., pp. 100–107, 2022.

[12] Y. Peng, S. Yan, and Z. Lu, "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets", in Proceedings of the 18th BioNLP Workshop and Shared Task, 2019, no. iv, pp. 58–65. doi: 10.18653/v1/W19-5006.

[13] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," in Chinese Computational Linguistics, 2019, pp. 194–206. doi: 10.1007/978-3-030-32381-3_16.

[14] HuggingFace, "Fine-tune a pretrained model", www.huggingface.com. https://huggingface.co/docs/transformers/en/training (accessed Feb. 05, 2024).

[15] E. Alsentzer et al., "Publicly Available Clinical BERT Embeddings", in Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78. doi: 10.18653/v1/W19-1909.

[16] Y. U. Gu et al., "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," ACM Trans. Comput. Healthc., vol. 3, no. 1, pp. 1–23, 2020, doi: 10.1145/3458754.

[17] A. Rogers, O. Kovaleva, and A. Rumshisky, "A Primer in BERTology: What We Know About How BERT Works", Trans. Assoc. Comput. Linguist., vol. 8, pp. 842–866, 2020, doi: 10.1162/tacl_a_00349.

[18] P. Vickers, L. Barrault, E. Monti, and N. Aletras, "We Need to Talk About Classification Evaluation Metrics in NLP," in Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, Nov. 2023, vol. 1, pp. 498–510. [Online]. Available: http://arxiv.org/abs/2401.03831 pp.498-510.

[19] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," in Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 2018, vol. 38, no. 3, pp. 353–355. doi: 10.18653/v1/W18-5446.

[20] G. Prasad, Y. Nie, M. Bansal, R. Jia, D. Kiela, and A. Williams, "To what extent do human explanations of model behavior align with actual model behavior?," in BlackboxNLP 2021 - Proceedings of the 4th BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, 2021, pp. 1–14. doi: 10.18653/v1/2021.blackboxnlp-1.1.

ChatGPT's Accuracy in Answering the National Medical Licensing Examination in Japan

Takayuki Nakano
Department of Pulmonary Medicine,
Graduate School of Medical Science
Kyoto Prefectural University of Medicine
Kyoto, Japan
tnakano@koto.kpu-m.ac.jp

***Abstract*—When applying generative AIs to the healthcare field, it is necessary to evaluate their performance. Although there is a previous study on English, we know little about Japanese. We evaluated ChatGPT's accuracy on the Japanese Medical Licensing Examination without modification of its sentence. ChatGPT (-4) achieved an accuracy that was good enough to pass the exam, as long as questions did not contain images. ChatGPT(-4) also showed its ability to make reasonable clinical inferences. While ChatGPT may have potential in healthcare use, we need to know more about its capabilities with respect to healthcare fields.**

Keywords-generative AI; ChatGPT; Japanese; Medical Licesing Examination.

I. INTRODUCTION

Large Language Models (LLMs), which are constructed from deep learning techniques on huge Web data sets, have made remarkable progress in recent years. In November 2022, Open AI Inc. launched ChatGPT. They fine-tuned the LLM for dialog-generating AI. To apply generative AI, such as ChatGPT, to the healthcare field, it is absolutely important to assess whether or not they have sufficient and correct medical knowledge. In addition, languages other than English are widely used in the healthcare field globally. There is a need to evaluate the competency of generative AI in languages other than English regarding healthcare affairs. In particular, Japanese is one of the hardest languages to master. Therefore, if a generative AI can demonstrate sufficient medical knowledge even in Japanese, it is reasonable to expect that it can do the same in many languages other than English or Japanese. These results would be a great motivation to apply generative AI to healthcare in many countries.

ChatGPT has already demonstrated its great ability to cover various areas, including the medicine and healthcare fields, even though it did not use a language model specific to healthcare. A previous study showed that ChatGPT (-3.5) had been able to pass the United States Medical Licensing Examination (USMLE) in all three categories [1]. According to the study, ChatGPT obtained accuracy equal to that of those who actually passed the examination. Indeed, a previous study showed that ChatGPT can pass the Japanese Medical Licensing Examination [2]. However, this report allowed translation from Japanese to English or modification of question sentences if they were not suited for ChatGPT in 2023, such as images. Therefore, little is known about ChatGPT's capabilities in the Japanese healthcare field.

The researcher assumed that ChatGPT could answer medical questions correctly even if they were posed only in Japanese. The aim of this study was to evaluate the accuracy of ChatGPT in the Japanese Medical Licensing Examination. In addition, we compared the scores between ChatGPT and the average scores of students who actually took the examination.

II. METHODS

In this section, we note about the Japanese Medical Licensing Examination and how to pose the question or its sentences and evaluate it.

A. Japanese Medical Licensing Examination

In Japan, the national examination for medical doctor candidates is held every year, and the actual posed questions and their correct answers are released on the website of the Ministry of Health, Labor and Welfare (MHLW. Almost all questions are written in Japanese. Examinees should select and answer from the presented choices; there is no descriptive question. The examination consists of three categories: general (e.g., Basic medicine such as anatomy, or public health), specific (e.g., Gastroenterology or cardiology), and content that must be mastered by a medical doctor. In addition, there are two main types of questions that could be solved with only knowledge (hereinafter referred to as "General question") and requiring clinical inference skills ('Clinical question'). The former gives the examinee one point per question if the answer is correct and the latter three points, and the latter is more similar to what medical doctors actually do. Candidates for the examination are required to exceed passing standards both in Contents have to be mastered and the others, and approximately 90% of all candidates pass the examination every year.

First, we collected all questions posed from 2018 to 2022 (N = 2000). Second, questions were excluded if they were classified as inappropriate by MHLW (N = 11) or contained images (e.g., photos of the patient, X-ray imaging, or figures) that ChatGPT cannot recognize in 2023 (N = 566). Overall, 1,423 questions were included in this analysis (Figure 1). For multiple-choice questions, a point was awarded only if all

choices were correct. In principle, no modification of the question sentence was allowed; however, only when the

Number of ChatGPT choices differed from the answer, one prompt was allowed to be re-presented. ChatGPT scores were evaluated for both GPT-3.5 and GPT-4.

B. Average number of medical students

In this study, we defined "medical student average" as the average score of students who actually took the examination. Because this group consists of both passed and failed students, the scores reflect the performance of students who have completed the education process for doctoral studies in Japan. The medical student average was calculated from the percentages of correct answers written in the books that explain the Japanese Medical Licensing Examination every year. The percentages were derived from questionnaires that more than 90% of all examinees had answered, so the data were reliable enough.

C. Statistical analysis

Fisher's exact test was used to evaluate significance. We calculated using EZR, a globally recognized software for analyzing medical statistics.

III. RESULTS

In this section, we describe the percentage of correct answers and scores of both the ChatGPT and medical student averages by question type.

A. Whole questions

When analyzing all questions, the accuracy rate of GPT-3.5 was 58.0% (826/1423) and that of GPT-4 was 84.0% (1196/1423). The scoring rate of GPT-3.5 was 59.7% (1080/1809), GPT-4 was 85.0% (1537/1809), and the medical student average was 85.6% (1548.574/1809), respectively (Table I and Figure 2). GPT-3.5 showed a much lower score than the medical student average; on the contrary, GPT-4 showed equal to the medical student average (without any significance).

B. By Questionare type

When calculated by questionnaire type, GPT-4 showed an ability similar to the medical student average (Table II and Figure 3). GPT-4 scores improved in almost all areas compared with GPT-3.5. Furthermore, although there was no significant difference, GPT-4 scored better than the medical student average on the Specifics, a category that included questions related to diseases, tests, or treatments.

We also examined scores separating general questions from clinical questions (Figure 4). While GPT-4 performed slightly inferior to the medical student average in the Clinical question, which requires clinical inference skills, GPT-4 was superior to the medical student average on the General questions, which focus on medical knowledge. There were no significant differences in either case.

TABLE I. WHOLE QUESTIONS AS ANALYZED

	respondent		
	GPT-3.5	***GPT-4***	***Medical student average***
Number of correct answers	826	1196	N/A
Percentage of Correct answers (%)	58.0 (826/1423)	84.0 (1196/1423	N/A
Total score	1080	1537	1548.574
Scoring rate (%)	59.7 (1080/1809)	85.0 (1537/1809)	85.6 (1548.574/1809)
p value	<0.001	0.587	Ref.
Odds ratio	0.249	0.948	Ref.
(95% CI)	(0.211-0.293)	(0.786-1.145)	Ref.
Average time taken to	8.78 (1-57)	3.03 (1-93)	N/A

N/A. Not available, Ref. Reference.

TABLE II. BY QUESTIONNAIRE TYPE

Questionare type	respondent		
	GPT-3.5	***GPT-4***	***Medical student average***
General	57.3 (323/564)	80.5 (454/564)	82.2 (463.641/564)
Odds ratio (95% CI)	0.289 (0.217-0.385)	0.889 (0.651-1.214)	Ref.
Specifics	52.1 (222/426)	83.8 (357/426)	80.0 (340.766/426)
Odds ratio (95% CI)	0.272 (0.200-0.372)	1.289 (0.894-1.862)	Ref.
Content must be mastered	65.3 (535/819)	88.6 (726/819)	90.9 (744.167/819)
Odds ratio (95% CI)	0.190 (0.142-0.252)	0.787 (0.563-1.098)	Ref.
General question	58.6 (407/694)	85.0 (590/694)	83.7 (580.796/694)
Odds ratio (95% CI)	0.276 (0.212-0.357)	1.10 (0.817-1.490)	Ref.
Clinical question	60.4 (673/1115)	84.9 (947/1115)	86.8 (967.778/1115)
Odds ratio (95% CI)	0.231 (0.186-0.287)	0.856 (0.669-1.094)	Ref.

Ref. Reference.

IV. DISCUSSION

We assessed how much ChatGPT has knowledge about healthcare in Japanese sentences. Our research has shown that ChatGPT (-4) might have sufficient knowledge equal to that of medical doctor candidates. Moreover, ChatGPT (-4) could make clinical inferences only in Japanese and was almost as accurate as medical students who had graduated from medical school.

Conventionally, ChatGPT is less accurate in its products on non-English prompts. Indeed, generative AI is very useful, but this situation will prevent its application to the healthcare field outside English-speaking countries, such as Japan. We assessed the ChatGPT's medical knowledge and clinical inference skills in Japanese sentences. We considered the

Medical Licensing Examination the most appropriate. First, the quality of the questions is guaranteed by a national institution. Second, it requires broad knowledge from basic medicine to internal medicine or surgery. Third, there was an ideal control group, the medical student average. There are also many medical specialist examinations in Japan, but most of them do not release actual questions or answers.

This study design was stricter than that of a previous report [2]. We had not allowed almost all modifications of the sentence, except for the number of choices. This fineness is partly used to evaluate the ability of ChatGPT in Japanese and to ensure that the questions are solved as closely as possible. Although there is skepticism about evaluating the significance of the results of generative AI, we believe that this rigorous study design allowed us to calculate significance.

We showed that ChatGPT (-4) can pass the Japanese Medical Licensing Examination if it excludes questions with images. Given the technical principles of generative AI, we could have presumed that it would perform better on questions requiring knowledge, but the fact that ChatGPT (-4) performed as well as the medical student average on questions requiring clinical inference skills was noteworthy. Inference skills are important in clinical practice and often take time for human medical students to master.

We have some limitations. First, we excluded more than a quarter of all questions, and most of the questions excluded contained images. Multimodal questions involving images may also be difficult for generative AI as human candidates. We consider that we can overcome this situation by collecting more than one thousand questions, and we hope that image recognition AIs will show us their desired performance. Second, this study specializes in medical doctor examinations. In addition to medical doctors, many other professions are involved in the healthcare field, and their contributions are significant. Therefore, you cannot simply judge that a generative AI can be applied to the healthcare field with only this study. However, it may be a milestone for the medical application of generative AI because knowledge of diseases and the ability to make clinical inferences are the basis for decision making in all healthcare fields.

ACKNOWLEDGMENT

I would like to thank Open AI Inc. and all its contributors. Furthermore, I am grateful to the reviewers for their useful suggestions.

REFERENCES

[1] T. H. Kung and M. Cheatham, “Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models.” PLOS Digit Health, vol. 2, e0000198. Feb. 2023, doi: 10.1371/journal.pdig.0000198.

[2] Y Tanaka and A. Nomura, “Performance of Generative Pretrained Transformer on the National Medical Licensing Examination in Japan.” PLOS Digit Health, vol. 3, e0000433. Jan. 2024, doi: 10.1371/journal.pdig.0000433.

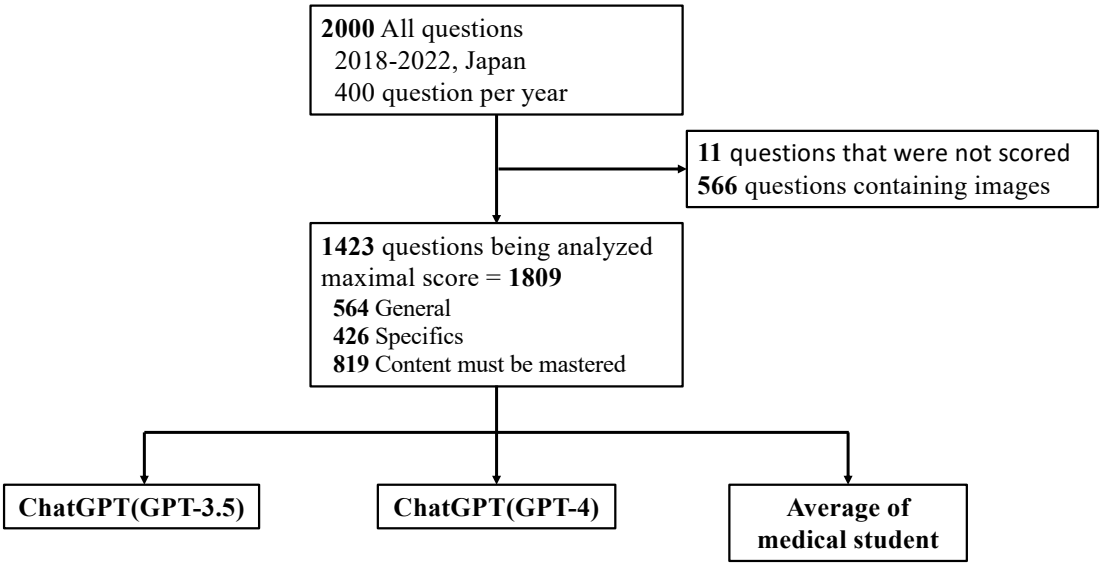


Figure 1. Questions regarding the inclusion and exclusion criteria of this study.

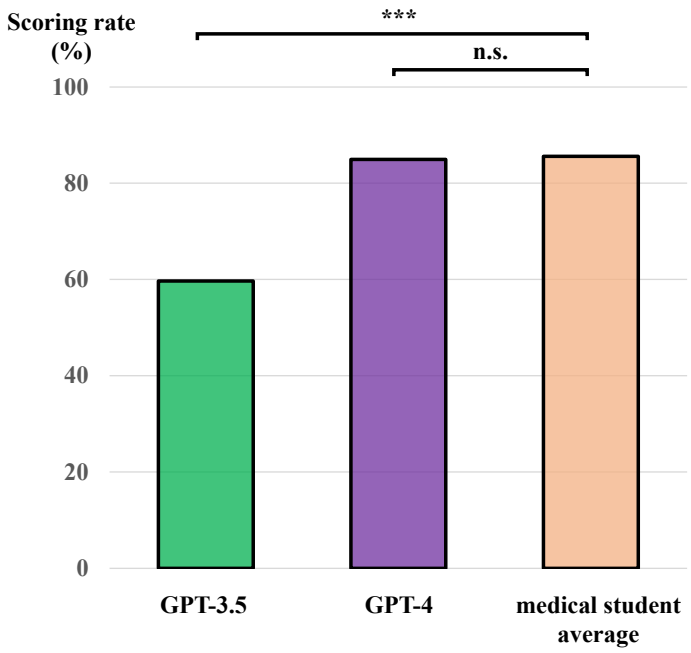


Figure 2. Total score in all questions. GPT-4 was superior to GPT-3.5 and equal to the medical student average.

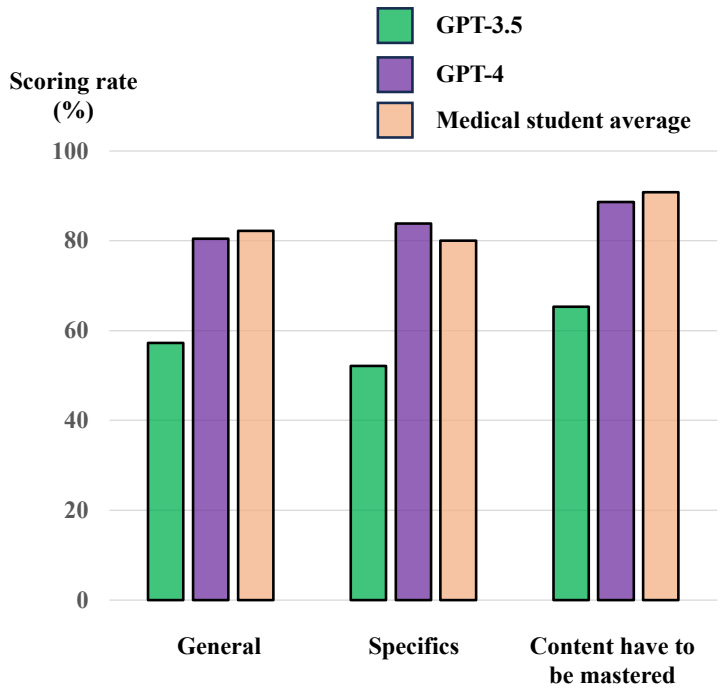


Figure 3. Score per category: General, Specific, and Content must be mastered. GPT-4 was generally similar to the medical student average. In particular, although there was no significance, ChatGPT (-4) was superior to the medical student average in Specific.

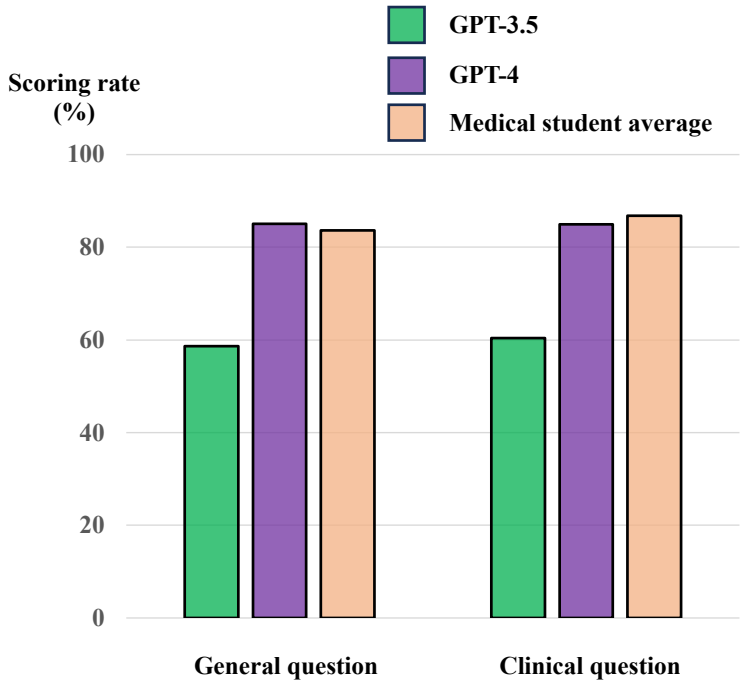


Figure 4. Score per category: General and Clinical questions. GPT-4 was generally similar to the medical student average. In particular, although there was no significant difference, ChatGPT (-4) was superior to the medical student average in the general question.

Priming Large Language Models for Personalized Healthcare

Madhurima Vardhan
Argonne Leadership Computing Facility
Argonne National Laboratory
Lemont, USA
0000-0003-4019-7832

Deepak Nathani
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, USA

Swarnima Vardhan
Department of Internal Medicine
Yale New Haven Health, Bridgeport Hospital
Bridgeport, USA

Abhinav Aggarwal
Department of Internal Medicine
Yale New Haven Health, Bridgeport Hospital
Bridgeport, USA

***Abstract*—Large Language Models (LLMs) have captured attention of researchers across different scientific fields. However, sensitive data access issues, model retraining, long compute time and lack of real-time results have limited the direct application of LLMs in fields such as healthcare fitness. Healthcare fitness is ripe to take advantage of the near-human efficiency and accuracy of LLMs due to ever increasing gap between human coaches and population that requires fitness coaching. In this work, we introduce a lightweight approach, priming LLMs, to develop an automated health coach that relies upon fundamental theories of behavior science and taps into the enormous potential of LLMs. We found that sentence length and conversation length were higher in primed LLMs compared to naïve context aware LLMs. Subsequently, we conducted a qualitative reviewer evaluation and report that the primed architectures were overall more appropriate and demonstrated higher empathy.**

***Index Terms*—Large Language Models, Personalized healthcare, fitness coaching, prompt engineering**

I. Introduction

Automated and personalized health coach assistants have the potential to reduce the cost of fitness and need for trained coaches who are required to cater the ever-increasing population suffering from non-communicable diseases and the rampant sedentary lifestyles [1] [2]. Given the rise of interest in personal health monitoring systems and increasing disparity with respect to the number of trained coaches versus the number of people who require fitness coaching, Large Language Models (LLMs) can be offered as an attractive solution to function as an automated health assistant. LLMs offer flexibility in performing a series of generalized tasks with near-human efficiency and accuracy [3]. LLMs such as GPT-3, have shown promise for task-oriented dialogue across a range of domains [3]. Both LaMDA and GPT-3 use the Transformer-based neural language models specialized for dialog applications [3]. In this work, we introduce a lightweight approach that constrains generalized LLMs to the specific task of functioning as a fitness coach and relies on established behavioral science models to enable empathetic and personalized conversations under different coaching scenarios.

While adapting or post-training LLMs using an unlabeled domain corpus has the potential to improve performance for end-tasks in a particular domain, the limitations around access to healthcare and personal data impede the application of LLMs for developing a personalized automated conversational assistant for fitness coaching [4] [5]. Thus, the use of LLMs in exercise coaching conversations remains relatively under-explored. Yet another reason for the lack of real-world automated fitness assistants using LLMs is also in part due to the complexity associated with health behavior change [6]. The field of behavior science has developed numerous frameworks for analyzing and influencing user behaviors, which has been critical in the design of personalized nudging programs in healthcare and fitness [7]. One such model is the Fogg's Behavior Model (FBM) that asserts the target behavior change of user can be explained across three axes by assessing: (1) motivation – is the user sufficiently motivated (2) ability – does the user have the ability to perform the given task, and (3) propensity – can the user plan or be triggered to perform the target activity. Several automated health assistant frameworks using different machine learning models have relied on the application of FBM to target behavior change, specifically for fitness coaching [8].

In this work, we explore how behavior science models, such as the FBM, could be infused into an LLM, and be used to constrain and/or guide the coaching conversations in a way that is consistent with established practice of human coaches. Towards this end, we propose priming LLMs as a lightweight approach that does not require additional model retraining and therefore precludes the need for any prior coach-user conversations. Priming essentially comprises of prompt engineering and design that can allow the model to be constrained for a specific task and in return has a higher probability to generate a more appropriate, favorable, and contextual conversation [9]. We encapsulate the FBM by priming the LLMs with example coach responses, mapped to the motivation, ability, and propensity of a user. Subsequently,

we qualitatively assess the conversations by independent human reviewers (n=5) generated by primed and unprimed LLMs across a commonly used LLM architectures, GPT-3. We found that sentence length, conversation length was higher in primed LLMs. Reviewers found that the primed architectures were overall more appropriate, showed higher motivation and greater empathy. Together, these experiments serve as a proof of concept of how LLMs and behavior science might be integrated, laying the foundation for future work around knowledge infusion in these conversational agents.

In the remainder sections of this work, Section II describes the overall approach and in Section III we discuss preliminary results, conclusion and future work.

II. Methods

In this section we describe our approach of priming LLMs with the FBM. ***Incorporating behavior science in LLMs by priming.*** We created a repository consisting of coach responses to different user scenarios by consulting expert behavior scientists and trained fitness coaches. The curated coach responses were tailored to a specific coach action across the three axes of FBM. For example, in the context of motivation, appropriate coach responses were created for encouragement, fun/temptation bundling, congratulating, and exemplifying core values/perceived benefit. Consequently, for ascertaining user ability, coach responses were constructed around providing educational information, barrier conversations, and recovery. Also, for propensity, appropriate examples were created for having goal conversations and activity planning. Using these examples, we primed GPT-3, and we refer to this approach as Behavior Science-based priming. We set the following model parameters for the open-source GPT-3 model, temperature (controls model randomness) to 0.9, maximum token length of 1024, top P (controls response diversity for likelihood responses) to 1 and frequency penalty (probability of verbatim model responses) to 0.9 and presence penalty (controls likelihood for new topics) to 0.6. We measure the quality of coach responses of the behavior science (BS) primed model, by comparing to naïve context-aware LLM model. For both the models (naïve and BS-primed), we emulated 10 different user scenarios exemplifying the need to elicit coach actions for sensing and boosting motivation, ability, and propensity of a user in a real-world scenario. We qualitatively evaluated the coach responses from both BS-primed and naïve LLMs by asking independent reviewers (n=5) to rate the conversations along different conversation dimensions, for example, coaching experience, empathy and appropriateness.

III. Conclusion and Future Work

Preliminary Results. Evaluating Behavior science (BS) primed LLM and naïve context aware LLM to function as an automated fitness coach. Qualitative analysis of coach-user conversations for both the BS-primed and naïve context aware, revealed that coach actions along the three FBM axes of motivation, ability and propensity were well-represented. However, we found that sentence length and conversation length were higher in BS-primed LLMs compared to naïve context aware LLMs. Furthermore, we qualitatively evaluated the quality of conversations of BS-primed and naïve context aware LLMs and found that all reviewers (n=5) preferred the BS-primed LLM responses with respect to coaching experience, empathy and appropriateness.

Conclusion and Future Work This is a proof of concept study of how fundamental models and medical knowledge can be used to encode healthcare information in LLM and enable them to function as an automated medical assistants for a more personalized experience without the need for any additional model retraining. User data such as obtained from wearable devices and smartphones can be used for automating the prompts via priming. Based on this framework, we will develop a zero shot learning approach for priming a LLMs that can function as an automated medical assistant for different clinical tasks, such that users will be able to directly chat with the LLM. Furthermore, we will quantitatively evaluate these tasks by human raters having domain expertise. We expect the ratings for the primed LLM to be significantly higher in terms of domain knowledge and empathy. As part of our future efforts, we will compare experimental results to real coaching assistant and automated virtual assistants.

IV. Acknowledgment

This research used resources of the Argonne Leadership Computing Facility, which is a U.S. Department of Energy Office of Science User Facility operated under contract DE-AC02-06CH11357.

References

[1] K. Matthias et al., "The mobile fitness coach: Towards individualized skill assessment using personalized mobile devices," Pervasive and Mobile Computing, vol. 9, no. 2, pp. 203-215, 2013.

[2] M. Vardhan et al.. "Walking with PACE-Personalized and Automated Coaching Engine." In Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, pp. 57-68. 2022.

[3] T. Brown et al., "Language models are few-shot learners." Advances in neural information processing systems 33: 1877-1901., 2020,

[4] Z. Ke et al., "Continual training of language models for few-shot learning." arXiv preprint arXiv:2210.05549, 2022.

[5] S. Yeom et al., "Privacy risk in machine learning: Analyzing the connection to overfitting." In 2018 IEEE 31st computer security foundations symposium (CSF), pp. 268-282. IEEE, 2018.

[6] S. Michie et al., "The Human Behaviour-Change Project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation." Implementation Science 12, no. 1: 1-12, 2017.

[7] B. J. Fogg, "A behavior model for persuasive design." In Proceedings of the 4th international Conference on Persuasive Technology, pp. 1-7. 2009.

[8] A. Lisowska et al., "From Personalized Timely Notification to Healthy Habit Formation: a Feasibility Study of Reinforcement Learning Approaches on Synthetic Data." In SMARTERCARE@ AI* IA, pp. 7-18. 2021.

[9] L. Reynolds et al., "Prompt programming for Large Language Models: Beyond the few-shot paradigm." In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1-7. 2021.

Efficacy of an AI-Based Weight Loss Digital Therapeutics Platform: A Multidisciplinary Perspective

Sarfraz Khokhar[1*], John Holden[2]
[1]Rasimo Systems, Raleigh, NC, USA
khokhar@rasimo.com
[2]Rockford College of Medicine, IL, USA
jholden@uwhealth.org

Catharine H. Toomer[3], Linda Whitby[4]
[3]Health Wellness and WL Centers, Aiken, SC, USA
wellness@total-weight-loss.com
[4]Rasimo Systems, Raleigh, NC, USA
whitby@rasimo.com

***Abstract*—Obesity is a global problem that has had a significant impact on society and the economy. The consequences are ominous with serious health risks. Millions of people are dying every year from complications of obesity and comorbidities. Despite efforts by governments and health agencies obesity continues to rise. Most of the approaches to management and treat obesity have not been successful because they did not shape people's lifestyle and the solutions that were provided for lifestyle modification are not multidisciplinary, they focus on only specific aspects. Obesity management mandates multidisciplinary approach with effective patient engagement, enhanced patient-healthcare provider communication, better adherence to therapy, minimize therapeutic inertia, motivation, more informed treatment decision by the healthcare provider, and addressing psychosocial conditions. We designed and developed an AI (artificial intelligence) based digital therapeutics platform to the multidisciplinary mandate for obesity management and treatment. We tested the efficacy of our proposed platform (solution) with a 24-week field trial and achieved 13.9% weight loss of the initial weight.**

Keywords- obesity; weight loss; digital therapeutics; artificial intelligence; expert systems.

I. INTRODUCTION

Obesity has matched epidemic proportions, with at least 2.88 million people dying every year as a result of being overweight or obese and a whopping economic and social impact of $1.7 trillion dollars [1]. The costs include $1.24 trillion in lost productivity and $480.7 billion in direct healthcare costs [2]. Once associated with high-income countries, obesity is now also prevalent in low and middle-income countries. Government agencies, non-governmental organizations, and the private sectors have been publishing their expert advice as good practices for a healthy lifestyle, in their research and field trials for decades and acknowledge that this pandemic is ever-increasing.

Despite ubiquitous information about nutrition and exercise, more fitness awareness, and more food and activity tracking devices, over 42% of the US adult population is living with obesity [3]. The world obesity rate grew proportionally as well [4]. The statistics show a significant increase from a decade ago, as depicted in Figure 1. The consequences are ominous; obesity is associated with serious health risks including heart, liver, gallbladder, kidneys, joints, breathing disorders, sleep apnea, diabetes, and several types of cancer [5]. The medical community continues struggling to find successful ways to encourage weight loss and provide effective interventions.

Lifestyle intervention faces challenges like compliance issues making weight loss difficult. Despite this, it continues to be a crucial component of obesity treatment. Digital tools augment lifestyle interventions by offering personalized support catering to the need for continuous interaction and support beyond conventional primary care settings. However, there is a need for a more comprehensive approach in utilizing digital tools to address the multifaceted aspects of obesity treatment effectively.

Traditional digital health methods of lifestyle modification have limited effectiveness in managing obesity as they lack multidisciplinary approach and engagement of HealthCare Provider (HCP) . The use of AI health coaching and predictive guidance for weight loss [6][7][8][9] is comparable with in-person HCP treatment, however it lacks patient engagement and treatment adherence.

Similarly, studies incorporating remote monitoring [10], motivational, moral, and community support [11][12], accountability [13], diet and nutrition management [14][15], physical activity tracking [16], and instant communication with the coaches through text messaging and video consultation [17][18] have been tried, however with limited success as they were monomodal.

Studies combining approaches and technologies showed better results. A clinical trial conducted showed that the use of a mobile application that used AI algorithms and gamification techniques to provide personalized feedback led to a significant reduction in body weight, body mass index (BMI), and waist circumference [19]. However, there is a need for effective, holistic, adaptive, cost effective, user-friendly, and integrated digital solution to manage obesity.
In the 21st century, AI and health technological advancement have enabled the development of digital therapeutics. Digital

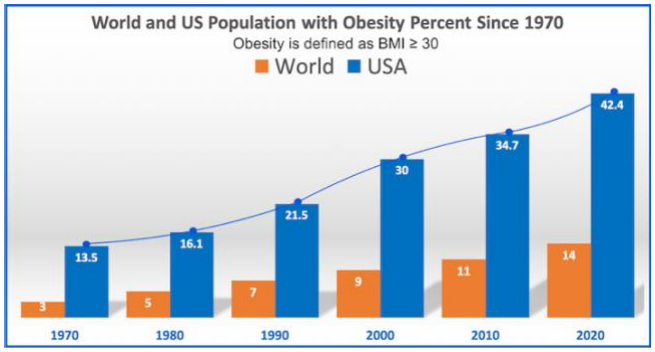


Figure 1: World and US obesity growth in the last six decades.

Therapeutics (DTx) are defined as evidence-based therapeutic interventions for patients by means of qualified software programs and medical devices to prevent, manage, or treat medical conditions.

Digital therapeutics can be more flexible than other treatment methods to address patients' individual needs [20]. These technologies employ various techniques, such as mobile applications, wearable devices, and online platforms, to improve the effectiveness of treatment interventions [21]. However, the current metabolic conditions such as obesity, diabetes, and cardiovascular diseases AI-based DTx would be an ideal complement to the pharmaceutical or even surgical weight loss offerings.

AI along with related technologies offer a promising approach for the management of obesity, as they use Machine Learning (ML) algorithms and/or expert systems (ES) to personalize treatment plans for patients.

We propose an AI-based DTx integrating all the approaches tried before but individually in a unified single platform, SureMediks. It includes short-term goals approach, tailored AI-based guidance and education on diet, nutrition, physical activities, and psychosocial conditions, effective and interactive patient-HCP communication, remote monitoring, motivation, accountability and community support.

SureMediks (our platform) includes an ES. Expert system is a branch of artificial intelligence (AI) that mimics the decision-making processes of human experts in specific domains. These systems are designed to provide guidance, advice, and recommendations to users based on their input and the knowledge (KB) rules programmed into the system [22]. These rules in KB can be updated as system learns new facts about the patients and their behavior. The integral components of an ES and its operation are depicted in Figure 3. The Knowledge Acquisition System of the ES extracts the expert knowledge and saves (learns) it in Knowledge Base (KB) as rules. Inference Engine (IE) activates these rules based on current and historical data and provides the guidance and education stored and learned in the KB. IE also updates the rules in KB dynamically. Explanatory Systems interprets patient's data and explain to the patients through charts and graphs in the mobile app.

In the context of patients' guidance and education, expert systems can provide personalized and interactive programs for managing and treating various health conditions, including obesity [23] and diabetes [24][25]. These systems can analyze patient data, such as medical history, symptoms, and lifestyle factors, and provide tailored recommendations and interventions to support patients in making informed decisions about their health.

AI feedback system was designed to address the primary barriers to successful weight loss, such as the complexity of dietary information, ineffective motivational strategies, and intermittent physical activity. By delivering real-time personalized feedback, SureMediks helps individuals remain on track, and offer corrective strategies when necessary. Additionally, it offers access to human expert guidance, which can further help individuals develop healthier behaviors that last longer.

The weight loss participants who reach their short-term goals have better long-term weight loss and ambitious short-term goals in the future[26] [27]. We used Khokhar WL Formula [28] to generate short-term goals, the formula is depicted in the equation below:

$$W_{loss} = \frac{\Delta W}{1-e^{-\frac{r\tau}{10}}}\left(e^{-\frac{rn}{10}} - e^{-\frac{r\tau}{10}}\right);\ r, \tau \neq 0; \quad \text{(Equation 1)}$$

Here, W_{loss}, ΔW, τ,and r are weight to lose at each short-term goal, total weight loss, time to lose weight in weeks, and r is a special parameter respectively, we called, r, the curve tension, n is the week number. For example, for n =1, 2, 3, it will determine the required weight loss for the first, second, and third weeks.

To assess the efficacy of our proposed AI-based DTx platform, SureMediks, we developed a prototype of the platform and set it up for a field trial. The implemented features and expert system's knowledge base were derived from a large research body and field trials mentioned previously in this section. In this paper we report summary of the filed trial and the results.

II. METHOD

This section describes our AI-based platform, SureMediks, field validation covering participants details, procedures and measurements.

A. Platform

Our platform consists of the following key elements: 1) An Internet-connected body composition scale to get patient's weight and related body metrics, 2) A mobile application through which patients receive tailored guidance, education, motivation, communicate with the HCP, interactive with accountability circle members for community support and visually can see the weight loss progress, 3) An AI agent acting as an expert system, and 4) A dashboard for the HCP to view patients' weight loss progress and interact with the patients.

B. Participants and weight loss goals

A participant sample of 1137 people of age 21 years and older from the USA, Canada, UK, and Australia were invited through emails and a weblink to participate in this field study. They were provided with key screening questions if they were determined and committed to losing weight that year, ready to be strictly focused on weight loss, ready and committed to be on a low-calorie trackable diet with daily trackable physical activity.

Finally, 391 participants took part in the trial from start to end. Of the 391 participants, 59% of the participants were female and 41% were male. Their education level, marital status, and other socioeconomic factors were not part of our selection criterion. However, their current weight, BMI, and age were among the primary concerns as we wanted to have diversity in age and weight buckets. Their start (baseline) mean weight, $\boldsymbol{\mu_{Start}}$, was 124.6 Kg with a standard deviation, $\boldsymbol{\sigma_{Start}}$, of 31.57 Kg, and a wide range of 65-181 Kg weight distribution. Mean age of the participants, $\boldsymbol{\mu_{Age}}$, was 43.56 years with a standard deviation, $\boldsymbol{\sigma_{Age}}$, of 12.60 years, and the range of 21-71 years. Their BMI mean, $\boldsymbol{\mu_{BMI}}$, was 43.9 Kg/m^2 with SD, $\boldsymbol{\sigma_{BMI}}$, of 8.5 Kg/m^2, 30 > BMI > 25 was considered overweight and BMI ≥ 30 was considered obesity as per World Health Organization (WHO) generic guidelines. The weight loss goal was 10% of the start weight however we set a stretch goal of 15% as the majority of the participants insisted on raising the bar.

C. Procedure and measures

The participants were provided with a WiFi-enabled smart body composition weighing scale and a mobile app, SureMediks. The study coordinators and coaches collaborated with the participants through a dashboard. The coaches, who were nutritionists, dietitians, and exercise instructors, had their own dashboards which they could log in and manage, communicate, and monitor the participants' progress, food intake, and physical activity. Figure 2 shows the high-level architecture of our implementation.

We created six groups of 391 participants based on their weight in six different weight buckets. Bucket 1 with participants of 65-85kg of weight, Bucket 2 for 86-105kg weight, Bucket 3 for 106-125kg, Bucket 4 for 126-145kg, Bucket 5 for 146-165kg, and Bucket 6 for the participants with the weight of 166-181kg. Our weighing scale maximum capacity was 181 Kg. These six weight buckets had 61, 78, 83, 60, 66, and 43 participants respectively, totaling 391 participants.

The participants in the study downloaded and installed the app on their smart devices and register their smart scale by scanning its ID or entering it manually. They provide their information, including age, height, preferred units, physical activity level, desired weight loss, duration, and group number. After signup, they were added to their coaches' dashboards. The baseline metrics were established automatically when they stepped on the scale for the first time, and weekly goals were sent by the intelligent agent based on the Khokhar WL formula [28]. The curve tension, r, adjusts dynamically based on weight loss performance, and participants are moved to a more suitable curve if they struggle to reach their weight loss goals.

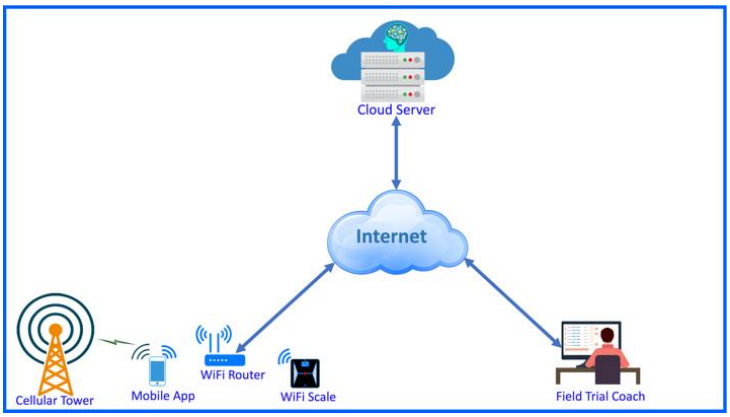


Figure 2: High-level implementation of the architecture: the participants have the scale and app, whereas the coaches have dashboards.

In this study, participants received feedback from an intelligent agent (ES) based on their current and historical data, each time they stepped on the scale along with education and guidance through the ES. The flow of ES is depicted in Figure 3. Two sample feedbacks are shown in Figure 4. Coaches also interacted with participants through text messages and video calls. The participants followed a low-calorie diet recommended by AI-based feedback mechanisms and the coaches, with food items shown in the app. Physical activity was chosen from a menu and tracked by AI and coaches. Participants were encouraged to step on the scale at least twice a week and could track their progress through charts in the app. Coaches focused on metabolic rates and weekly weight loss, providing additional guidelines if goals were not achieved. Participants formed accountability circles for support and motivation, and alerts were set up to notify if weight gain occurred. Participants were proactive in making corrections to their diet, physical activities, and lifestyle based on feedback and guidance from the ES.

SureMediks, encouraged participants to engage in challenges within their accountability circle, facilitated by the app. There were six challenges to lose 3% weight each, and the platform tracked the number of challenges participants took part in. In addition to community support, participants received daily motivational quotes selected by the AI agent based on their progress or challenges. After 26 weeks, participants' weekly weights were noted and their weight loss progress statistics were analyzed using MS Excel data analysis tools.

D. Results

The detailed weight loss statistics of each of the six buckets is as follows: For Weight Bucket1, 65-85 Kg, the

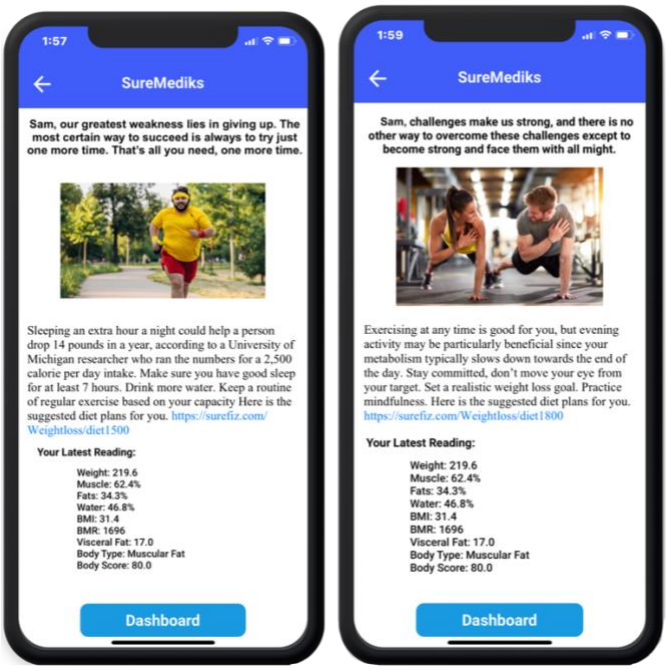


Figure 4: Samples of guidance from the ES, first part is motivational and second part is feedback and guidance.

mean weight loss, μ_{wl1}, was 10.1 kg, standard deviation, σ_{wl1} = 3.4 kg, mean weight loss percentage of 13.3, with a 95% confidence interval (CI) of 12.18% -14.38%, and BMI loss (drop) of 4.3 points. For Weight Bucket2, 86-105 Kg, the mean weight loss, μ_{wl2}, was 13.6 Kg, standard deviation, σ_{wl2} = 4.4 Kg, mean weight loss percentage of 14.2, with a 95% confidence interval (CI), 13.20% -15.19 %, and BMI loss (drop) of 5.2 points. For Weight Bucket3, 106-125 Kg, the mean weight loss, μ_{wl3}, was 15.9 Kg, standard deviation, σ_{wl3} = 5.2 Kg, mean weight loss percentage of 14.0, with a 95% confidence interval (CI), 13.03% - 14.96%, and BMI loss (drop) of 5.9 points. For Weight Bucket4, 126-145 Kg, the mean weight loss, μ_{wl4}, was 19.1 Kg, standard deviation, σ_{wl4} = 5.9 Kg, mean weight loss percentage of 14.5, with a 95% confidence interval (CI), 13.41% - 15.58%, and BMI loss (drop) of 6.8 points. For Weight Bucket5, 146-165 Kg, the mean weight loss, μ_{wl5}, was 19.4 Kg, standard deviation, σ_{wl5} = 6.8 K, mean weight loss percentage of 12.53, with a 95% confidence interval (CI), 11.45% - 13.60%, and BMI loss (drop) of 6.7 points. For Weight Bucket6, 166-181 Kg, the mean weight loss, μ_{wl6}, was 25.5 Kg, standard deviation, σ_{wl6} = 7.3 Kg, mean weight loss percentage of 14.8, with a 95% confidence interval (CI), 13.54% - 16.07% , and BMI loss (drop) of 8.6 points.

Overall, for all 391 participants, 65-181kg, the mean weight loss, μ_{wl}, 17.27 Kg, with standard deviation, σ_{wl6} = 7.0 Kg, mean weight loss percentage of 13.89, with a 95% confidence interval (CI), 13.45% – 14.35%, and BMI loss (drop) of 8.6 points. The p-value was significant, $p < 0.0001$, for all results, confidence interval (CI), 13.54% - 16.07% ,

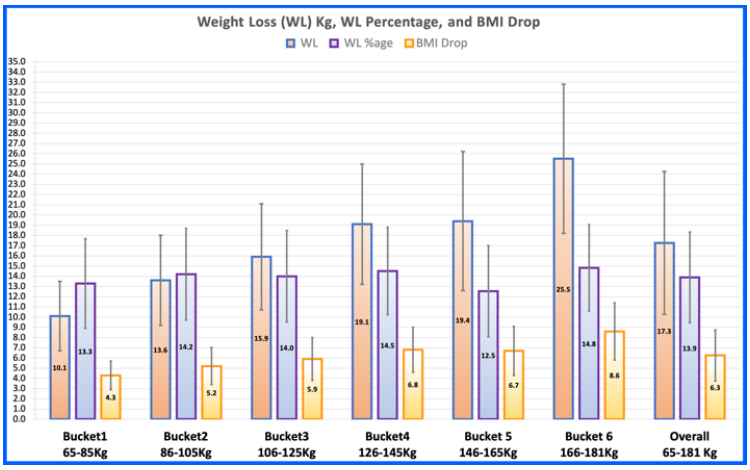


Figure 5: Higher BMI drop with larger weight buckets. the weight loss percentage is similar across all the buckets.

and BMI loss (drop) of 8.6 points. Figure 5 depicts the key results: mean and standard deviation of weight loss, weight loss percentage, and BMI loss (drop).

Figures 6 shows the weekly plotted mean weight loss progress in kilo grams of all the buckets combined (391 participants). In this plot, the Amber curve depicts the weekly weight loss progress for the period of the trial and the blue line shows the weekly predicted mean weight of the participant per the Khokhar Weight Loss formula (Equation 1). The predicted weight loss curve could serve as the trend curve as well.

III. DISCUSSION

This study suggests that digital platforms are efficient for weight loss programs. We found that participants had a mean weight loss of 13.9% from baseline using an AI-assisted lifestyle intervention only. The study set a stretch weight loss goal of 15% based on participants' preferences and determination which was found to play a vital role in weight loss efforts. Dividing the weight loss goal into smaller weekly goals made it less overwhelming and increased participants' sense of control and confidence. The study also found that AI guidance, extensive communication and guidance from coaches, motivation, accountability, and community support were driving factors in achieving these outstanding weight loss goals. The use of timely guidance and feedback, along with extensive communication, led to better outcomes. Motivation derived from internal and external factors, along with accountability and community support, played significant roles in participants' weight loss. Food journaling and physical activity tracking also contributed to healthier food choices and increased physical activity. Overall, a comprehensive approach with the optimal use of technology is effective for weight and obesity management.

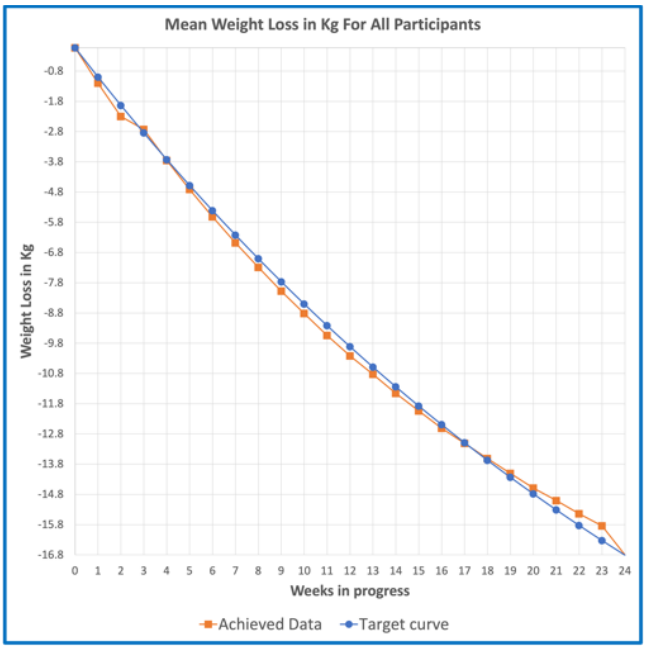


Figure 6: Weekly weight loss progress of all 391 participants. Average weekly weight loss was 0.71 Kg.

IV. CONCLUSION AND FUTURE WORK

Consistent weight loss needs a multidisciplinary approach. Determination, motivation, effective communication, diet, physical activity, accountability, and tailored guidance and education are vital elements. Digital therapeutics for obesity have the potential to significantly improve patient adherence and treatment outcomes and can deliver a framework where these key elements asynchronously and coherently work for the best patient-HCP engagement and optimal patient outcome. It is a promising way to address the global pandemic of obesity and warrants significant investment for further development. AI plays a vital role in delivering tailored guidance and education to the patient and catalyze the effectiveness of DTx. With a properly designed and operated digital therapeutics platform surpassing the benchmark of 10% weight loss in 24 weeks is feasible with an effective diet and physical plan along with the vital elements of a multidisciplinary approach, which a DTx platform can deliver effectively using ES.

Our future work is focused on studying how SureMediks can effectively complement medical weight loss with Glucagon-Like Peptide (GLP-1) and similar weight loss medication and post metabolic surgery weight loss.

ACKNOWLEDGMENT

We would like to thank our participants and their commitment to their dedicated time. Their compliance and dedications to our field trial were commendable. We would also like to thank Mr. Kannan Palaniswami for his efficient technical support and Ms. Helga Andersen for her moral support for completing the field trial and compilation of this paper.

REFERENCES

[1] World Health Organization. Obesity. (2021). https://www.who.int/news-room/facts-in-pictures/detail/6-facts-on-obesity [Accessed February 8, 2024].

[2] Milken Institute. Economic impact of excess weight now exceeds $1.7 trillion. Science Daily. (2018). https://www.sciencedaily.com/releases/2018/10/181030163458.htm [Accessed February 8, 2024].

[3] Overweight and Obesity Statistics. National Institute of Diabetes and Digestive and Kidney Diseases. (2020). https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity [Accessed February 8, 2024].

[4] C. Boutari, C.S, Mantzoros, "A 2022 update on the epidemiology of obesity and a call to action: as its twin COVID-19 pandemic appears to be receding, the obesity and dysmetabolism pandemic continues to rage on," Metabolism, vol. 133, pp. 155-217, 2022.

[5] G.A. Bray, et al., "The science of obesity management: An endocrine society scientific statement," Endocr Rev., vol. 39 no. 2, pp. 79-132, 2018.

[6] N. Stein, K.A, Brooks, "Fully automated conversational artificial intelligence for weight loss: longitudinal observational study among overweight and obese adults," JMIR Diabetes, vol. 2, no. 2, pp. e28, 2017

[7] E. M. et al., "Using artificial intelligence to optimize delivery of weight loss treatment: Protocol for an efficacy and cost-effectiveness trial." Contemp Clin Trials. Vol. 124, pp. 107-129, 2022.

[8] C.A. Pellegrini, A.F. Pfammatter, D.E. Conroy, B. Spring , "Smartphone applications to support weight loss: current perspectives," Adv Health Care Technol. vol. 1, pp. 13-22, 2015.

[9] R.A. Krukowski, J. Harvey-Berino, T. Ashikaga, C.S. Thomas, N. Micco, "Internet-based weight control: the relationship between web features and weight loss," Telemed J E Health., vol. 14, no. 8, pp. 775-782, 2008.

[10] L. Hu et al., "Challenges of conducting a remote behavioral weight loss study: Lessons learned and a practical guide," Contemp Clin Trials., vol. 108, pp. 106-122, 2021.

[11] S. Soini, P. Mustajoki, J. G. Eriksson, "Long-term weight maintenance after successful weight loss: Motivational factors, support, difficulties, and success factors," Am. J. Health Behav., vol. 42, no. 1, pp. 77-84, 2018.

[12] D.S. West, et al., "A motivation-focused weight loss maintenance program is an effective alternative to a skill-based approach," Int. J. Obes. (Lond)., vol. 35 no. 2, pp. 259-269, 2011.

[13] T.W. Bradford, S.A. Grier, G.R. Henderson, "Weight loss through virtual support communities: A role for identity-based motivation in public commitment," Journal of Interactive Marketing, vol. *40*, no. 1, pp. 9–23, 2017.

[14] J.F. Hollis et al., "Weight loss maintenance trial research group. weight loss during the intensive intervention phase of the weight-

loss maintenance trial," Am. J. Prev. Med., vol. 35, no. 2, pp. 118-26, 2008.
[15]A. McTiernan, "Exercise effect on weight and body fat in men and women," Obesity (Silver Spring), vol. 15, no. 6, pp. 1496-512, 2007.
[16] K. Frie, J. Hartmann-Boyce, S. Jebb, S., J. Oke, P. Aveyard, "Patterns in weight and physical activity tracking data preceding a stop in weight monitoring: observational analysis," J. Med. Internet Res., vol. 22, no. 3, pp. 157-190, 2020.
[17] L.E. Burke, J. Wang, M.A. Sevick, "Self-monitoring in weight loss: a systematic review of the literature," J. Am. Diet Assoc., vol. 111, no. 1, pp. 92-102, 2010.
[18] G. Turner-McGrievy, D. Tate, "Tweets, apps, and pods: results of the 6-month mobile pounds off digitally (mobile pod) randomized weight-loss intervention among adults," J. Med. Internet Res., vol. 13, no. 4, pp. 1-14, 2011.
[19] L. Hebden et al., "Mobile health intervention for weight management among young adults: A pilot randomized controlled trial," J. Hum. Nutr. Diet, vol. 27, no. 4, pp. 322-32, 2014.
[20] J.S. Hong, C. Wasden, D.H. Han, "Introduction of digital therapeutics," Comp. Methods Programs Biomed, vol. 209, pp. 106-319, 2021.
[21] N. Hinchliffe, M.S. Capehorn, M. Bewick, J. Feenie, "The potential role of digital health in obesity care," Adv. Ther. Vol. 39, no. 10, pp. 4397-4412, 2022.
[22] T.S. Sayed, "Application of expert systems or decision-making systems in the field of education," Information technology in industry, vol. 27, no. 3, pp. 1176-1183, 2021.
[23] C.L. Chi, W. Nick Street, J.G. Robinson, M.A. Crawford, "Individualized patient-centered lifestyle recommendations: an expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options.,"*Journal of biomedical informatics*, vol. *45*, no. 6, pp. 1164–1174, 2012.
[24] P. Valsalan, N. U. Hasan, U. Farooq, M. Zghaibeh, I. Baig, "IoT Based Expert System for Diabetes Diagnosis and Insulin Dosage Calculation. *Healthcare (Basel, Switzerland)*," vol. *11*, *no.*1, pp. 12-21, 2022.
[25] P. Jirků, M. Andel, P. Hájek, P. Expertní," Systém v diagnostice diabetu [An expert system for the diagnosis of diabetes]," *Casopis lekaru ceskych*, vol. *125*, no. 2, pp. 49–52, 1986.
[26] R.W. Jeffery, R.R. Wing, R.R., Mayer, "Are smaller weight losses or more achievable weight loss goals better in the long term for obese patients?" J. Consult. Clin. Psychol., vol. 66, no. 4, pp. 641-645, 1998.
[27] Mayo Clinic. The Mayo Clinic Diet. A weight-loss program for life. https://www.mayoclinic.org/healthy-lifestyle/weight-loss/in-depth/mayo-clinic-diet/art-20045460 [[Accessed February 8, 2024].
[28] Khokhar, S. (2022). Methods and systems for interactive weight management**.** U.S. Patent No 11,353,358. Washington, DC: U.S. Patent and Trademark Office.

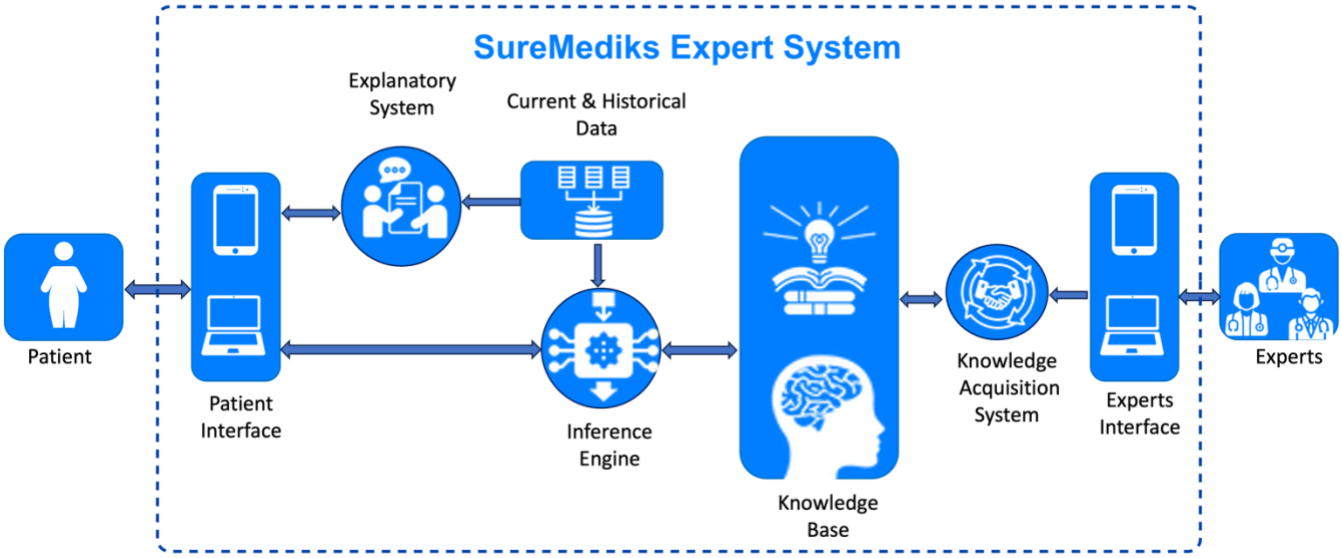


Figure3: Distributed Architecture and the operation flow of our SureMediks Expert System.