ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ISSN: 3041-0894 (Print)
ISSN: 3029-2387 (Online)
Volume 1 · Issue 2
April 2024

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

Print ISSN: 3041-0894
Online ISSN: 3029-2387

Artificial Intelligence in Health aims to provide a freely accessible multidisciplinary and comprehensive platform for researchers, scientists, and AI in health and medicine sciences practitioners to publish and exchange cutting-edge advancements, insights, technological development and innovations at the intersection of artificial intelligence (AI) and health. The journal seeks to explore the transformative potential of AI in improving and understanding health and medicine research outcomes, enhancing clinical decision-making, optimizing resource allocation, and addressing various challenges in the multidisciplinary field of health.

About the Publisher

AccScience Publishing is a publishing company based in Singapore. We publish a range of high-quality, open-access, peer-reviewed journals and books from a broad spectrum of disciplines.

Contact Us

Managing Editor
aih.office@accscience.sg

AccScience Publishing
8 Burn Road, #15-03 Trivex, Singapore 369977.

Volume 1 • Issue 2 • April 2024
ISSN 3041-0894 (print) ISSN 3029-2387 (online)

ARTIFICIAL INTELLIGENCE IN HEALTH

ACCSCIENCE PUBLISHING

Access Science Without Barriers

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ISSN: 3041-0894 (print)
ISSN: 3029-2387 (online)

Editorial and Production Credits

Publisher: AccScience Publishing
Managing Editor: Irene Zhao
Production Editor: Sharmila Velapasamy
Article Layout and Typeset: Sinjore Technologies (India)

Supplementary file

Supplementary files of articles can be obtained at https://accscience.com/journal/AIH/1/2.

Disclaimer

Artificial Intelligence in Health

Editorial Board

ACCSCIENCE PUBLISHING

Volume 1 | Issue 2 | April 2024 | Page 1 – 143 |

Artificial Intelligence in Health

https://accscience.com/journal/AIH

CONTENTS

REVIEW ARTICLES

1 **Artificial intelligence in the battle against COVID-19: A comprehensive review**
Emma Yann Zhang, Adrian David Cheok, Zhigeng Pan, Jun Cai, Ying Yan

16 **LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties**
Ummara Mumtaz, Awais Ahmed, Summaya Mumtaz

29 **Enhancing health-care security: The role of blockchain and consensus mechanisms**
Asmaul Hosna, Nujhat Tabassum Rahman, Supriya Dewanjee, Zulfikar Alom, Elmustafa Sayed Ali, Mohammad Abdul Azim, Rashid A. Saeed

48 **Artificial intelligence in diagnosis and monitoring of atopic dermatitis: From pixels to predictions**
Pratheek Jain, Farhan Zameer, Kounaina Khan, Vinay Alva, Ravish Huchegowda, Ali Jawad Akki, Raghu Anjanapura Venkataramanaiah, Muthuchelian Krishnasamy, Dilip Apturkar, Raghavendra Hallur Laxmanashetty

66 **The perspectives of eye care professionals on the integration of artificial intelligence in eye care practices: A systematic review**
Obehi Suzan Idogen

ORIGINAL RESEARCH ARTICLES

76 **Discovery of new antibiotics using AI-guided spectroscopy and 3D drug-protein computer simulation technologies to combat MDR bacteria-associated mortality**
Asit Kumar Chakraborty, Meghna Maity, Sumana Sahoo

96 **Efficient schema-less text-to-SQL conversion using large language models**
Youssef Mellah, Veysel Kocaman, Hasham Ul Haq, David Talby

107 **Development and analysis of medical instruction-tuning for Japanese large language models**
Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, Satoshi Kodera

117 **Factors associated with social determinants of health mentions in PubMed clinical case reports from 1975 to 2022: A natural language processing analysis**
Julio Bonis, Veysel Kocaman, David Talby

132 **Enhancing patient safety through integrated sensor technology and machine learning for bed-based patient movement detection in inpatient care**
Jonathan Mayer, Rejath Jose, Molly Bekbolatova, Chris Coletti, Timothy Devine, Milan Toma

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

Artificial intelligence in the battle against COVID-19: A comprehensive review

Emma Yann Zhang[1†*], Adrian David Cheok[2†*], Zhigeng Pan[1], Jun Cai[2,3], and Ying Yan[2]

[1]School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, Jiangsu, China

[2]School of Automation, Nanjing University of Information Science and Technology, Nanjing, Jiangsu, China

[3]Anhui Jianzhu University, Shushan District, Hefei City, Anhui, China

†*These authors contributed equally to this work.*

***Corresponding authors:**
Emma Yann Zhang
(202351620003@nuist.edu.cn)
Adrian David Cheok
(adrian@imagineeringinstitute.org)

Citation: Zhang EY, Cheok AD, Pan Z, Cai J, Yan Y. Artificial intelligence in the battle against COVID-19: A comprehensive review. *Artif Intell Health.* 2024;1(2): 1-15. doi: 10.36922/aih.2401

Received: December 11, 2023

Accepted: January 15, 2024

Published Online: April 4, 2024

Abstract

The COVID-19 pandemic has precipitated a global crisis, affecting all facets of human life. The rapid spread of the virus necessitated urgent responses from the healthcare sector, with artificial intelligence (AI) taking center stage as a pivotal tool in this fight. This paper provides a comprehensive review of the multifaceted role of AI during the pandemic, spanning from early detection and diagnosis to treatment, management, and the development of vaccines. We delve into the ethical and societal implications of deploying AI in such critical scenarios, discussing data privacy, algorithmic bias, and accessibility. The paper also presents various case studies, highlighting country-specific implementations and the dichotomy of success stories and failures. Furthermore, we explore the future directions of AI in healthcare, emphasizing emerging technologies and policy recommendations that could shape post-pandemic health-care systems. The conclusion synthesizes these insights, reflecting on the lessons learned and the prospective landscape of AI in global health. This paper aims to serve as a cornerstone for policymakers, health-care providers, and AI researchers, guiding the responsible and effective integration of AI in future health-care strategies.

Keywords: COVID-19; Artificial intelligence; Machine learning; Data privacy; Wearable technologies: Telemedicine; Vaccine development; Ethical implications

1. Introduction

1.1. Background

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, has had a profound impact on global health, economies, and daily life.[1-3] First identified in Wuhan, China, in late 2019, the virus has since spread worldwide, leading to significant morbidity and mortality.[4] Governments and healthcare systems have been stretched to their limits, endeavoring to manage the crisis effectively.[5]

In this global crisis, artificial intelligence (AI) has emerged as a powerful tool in the fight against COVID-19. AI technologies have been employed in various capacities, from early detection and diagnosis to vaccine development and public health policy planning.[6] The adaptability and computational power of AI have made it a valuable asset in rapidly evolving scenarios, where timely and data-driven decisions are crucial.

This paper aims to provide a comprehensive review of the multifaceted role of AI in combating the COVID-19 pandemic. We will explore its applications in detection, treatment, analytics, and ethical considerations, among other aspects.

1.2. Rationale

The COVID-19 pandemic has presented unprecedented challenges to global health,[5] necessitating rapid and innovative solutions. AI has been at the forefront of these solutions, offering a range of applications from diagnostic tools to predictive modeling. The rationale for studying AI in this context is multifaceted.

AI technologies have demonstrated their potential in enhancing the efficiency and accuracy of diagnostic procedures, contributing to faster and more reliable detection of COVID-19 cases.[7,8] Furthermore, AI has played a crucial role in analyzing vast amounts of data to identify patterns and make predictions, aiding in the development of strategies to mitigate the spread of the virus.[9]

The integration of AI in managing the pandemic has also extended to the realm of public health and policy planning. AI-driven tools have been utilized to combat misinformation on social media, ensuring that accurate and reliable information is disseminated to the public.[10] In addition, AI has been employed to analyze the psychological impact of the pandemic on health-care professionals, providing valuable insights into the mental health challenges faced by frontline workers.[11,12]

The study of AI in the context of the COVID-19 pandemic is not only crucial for addressing the immediate challenges posed by the virus but also holds significant implications for future pandemic preparedness and response. By understanding the capabilities and limitations of AI in this context, we can pave the way for more robust and resilient health-care systems, better equipped to handle future global health crises.[13]

1.3. Objectives

The primary objectives of studying AI in the context of the COVID-19 pandemic are multifaceted, aiming to harness the potential of AI in various domains to combat the challenges posed by the pandemic. The objectives are listed as follows:

(i) Enhancing diagnostic and prognostic accuracy: AI has played a crucial role in improving the accuracy of COVID-19 diagnosis and prognosis. Machine learning models have been developed to interpret clinical, laboratory, and imaging data, aiding health-care professionals in making more informed decisions.

(ii) Optimizing resource allocation: The pandemic has put an unprecedented strain on healthcare systems worldwide, necessitating optimal resource allocation. AI has been instrumental in managing hospital resources, including the distribution of medical supplies and the allocation of hospital beds.

(iii) Supporting mental health: The mental health implications of the pandemic are profound, with health-care workers being particularly affected. AI-powered tools have been developed to provide mental health support and resources, aiding in the mitigation of the psychological impact of the pandemic.

(iv) Facilitating remote learning and work: The shift to remote learning and work has been one of the most significant changes during the pandemic. AI has played a role in enhancing the effectiveness of remote learning and work environments, ensuring continuity in education and professional activities.

(v) Aiding in vaccine development and distribution: AI has been pivotal in accelerating the development and distribution of COVID-19 vaccines. Machine learning models have been utilized to analyze vast datasets, aiding in the identification of potential vaccine candidates and optimizing distribution logistics.

(vi) Improving public health surveillance: Enhancing public health surveillance has been a key objective in the fight against COVID-19. AI has been employed to analyze data from various sources, providing real-time insights into the spread of the virus and informing public health interventions.

(vii) Addressing misinformation: The pandemic has been accompanied by an infodemic of misinformation. AI has been utilized to identify and counteract misinformation, ensuring that accurate and reliable information is disseminated to the public.

(viii) Promoting equity and inclusion: Ensuring equity and inclusion in the response to the pandemic is paramount. AI has the potential to identify and address disparities in healthcare access and outcomes, promoting a more equitable response to the pandemic.

2. Scope and limitations

The COVID-19 pandemic has precipitated an unprecedented reliance on AI across various domains of healthcare and public health management. This review endeavors to delineate the multifaceted applications of AI during the COVID-19 crisis, encompassing disease surveillance, diagnostic methodologies, therapeutic development, and the optimization of patient care protocols. A particular emphasis is placed on the pivotal role of AI in enhancing the efficacy of diagnostic algorithms, which have been instrumental in the identification and management of COVID-19 cases. Furthermore, the review will scrutinize the ethical dimensions and data privacy considerations that are intrinsically linked to the utilization of AI technologies in the milieu of public health emergencies.

The significance of AI in the healthcare domain during the COVID-19 pandemic has been extensively documented, with particular regard to its future potential and current applications.[6,14] Moreover, the motivations and imperatives for leveraging AI and big data in response to the COVID-19 crisis have been thoroughly explored in the literature.[15] An early review has also highlighted the contributions and current constraints of AI in combating COVID-19.[16] This review builds on the foundational work of previous studies but extends beyond them by offering a more comprehensive, ethically informed, and future-oriented analysis of AI in the context of the COVID-19 pandemic.

Notwithstanding the extensive scope of this review, it is imperative to acknowledge the inherent limitations that circumscribe its breadth. The dynamic and rapidly evolving landscape of AI technology, coupled with the continuous emergence of novel research, inherently limits the capacity to encapsulate all current initiatives within the confines of this paper. In light of the voluminous literature pertaining to AI and COVID-19, the focus will be primarily directed towards peer-reviewed articles and seminal case studies, excluding non-peer-reviewed "grey literature" and unpublished research work. In addition, the time constraints inherent to the writing process may prevent the inclusion of the most recent developments in the field.

In recognition of these limitations, this review does not claim to be exhaustive; rather, it seeks to furnish a comprehensive and representative overview of the current state of AI applications within the context of the COVID-19 pandemic, thereby providing a foundational understanding of the subject matter and a platform for future research endeavors.

2.1. Ethical considerations

The ethical implications of AI deployment in healthcare, especially during a pandemic, are profound and multifarious. Issues pertaining to data privacy, informed consent, and the potential for algorithmic bias necessitate careful consideration, particularly in the context of public health and the management of personal medical data.

2.2. Technological constraints

The technological constraints that define the scope of this review are equally significant. While AI holds significant potential to enhance pandemic response strategies, its effectiveness depends on the availability of high-quality data, the robustness of algorithms, and the strength of the underlying infrastructure that implements the solutions.[17]

3. Organization of the paper

This review is structured to facilitate a comprehensive understanding of the multifarious applications of AI in the context of the COVID-19 pandemic. The sections are systematically organized to provide a logical progression from historical precedents to future predictions, encompassing the entire spectrum of AI's contributions to pandemic management.

Section 4 delineates the rigorous approach employed in gathering existing literature. It details the strategies used in the literature search, the inclusion and exclusion criteria, and the methods of analysis adopted to synthesize the information.

Section 5 explores the historical development of AI in healthcare, with particular emphasis on its role in disease detection and diagnosis, vaccine development, treatment strategies, and epidemiology modeling. This section lays the groundwork for understanding AI's application in the COVID-19 pandemic.

Section 6 explores the technical aspects of AI in detecting and diagnosing COVID-19. It is further broken down to highlight the specific contributions of imaging techniques, natural language processing (NLP), and wearable technologies.

Section 7 examines AI's critical role in drug discovery, patient management, and the evolving realm of telemedicine. It underscores AI's transformative impact on improving patient care and optimizing healthcare services.

Section 8 investigates AI's predictive capabilities in epidemiological modeling, resource distribution, and social media analysis for public sentiment and reaction to the pandemic.

Section 9 addresses the ethical dilemmas and societal implications of employing AI during a healthcare crisis. It focuses on crucial issues such as data privacy, algorithmic bias, and unequal access to AI technologies.

Section 10 presents a series of case studies that demonstrate AI's practical applications across different global and sociopolitical settings. It offers a critical evaluation of both successful and less successful AI implementations.

Section 11 looks forward to emerging technologies that may influence the future role of AI in pandemic response. It provides policy recommendations to maximize the benefits of AI in this context.

4. Methodology

This comprehensive review employs a meticulous and expansive literature search strategy designed to encompass the full spectrum of AI applications in the context of the COVID-19 pandemic. This strategy ensures the inclusion of a diverse array of studies that provide a representative cross-section of the current state of knowledge.

4.1. Literature search strategy

The development of our search criteria was a collaborative and iterative process, involving a consensus among a team of interdisciplinary researchers. A comprehensive search was conducted across multiple academic databases and search engines, including PubMed, Scopus, Web of Science, and Google Scholar, to ensure a thorough survey of the existing literature. The search strategy was augmented using Boolean operators, truncation, and wildcard characters to maximize the retrieval of relevant studies.

The search was intentionally broadened to include studies from a multitude of disciplines, recognizing the inherently interdisciplinary nature of AI applications in pandemic response. This approach facilitated the inclusion of research spanning the domains of healthcare, public health, computer science, and social sciences.

The temporal scope of the search was defined to include studies published from the start of the pandemic in late 2019 through to the present day. The search strategy was periodically updated to incorporate the latest research findings, ensuring the review is up-to-date.

A carefully curated list of keywords and topic headings was employed, encompassing terms such as "COVID-19," "SARS-CoV-2," "artificial intelligence," "AI," "machine learning," "deep learning," "neural network," "pandemic," "public health," and "telemedicine," among others. This strategy was instrumental in unearthing studies that specifically addressed the multifaceted applications of AI in the pandemic milieu.

4.2. Inclusion and exclusion criteria

The integrity of this review is subject to a stringent set of inclusion and exclusion criteria, meticulously crafted to ensure the selection of studies that provide robust and relevant insights into the applications of AI during the COVID-19 pandemic. These criteria serve as a safeguard against methodological inconsistencies and form the foundation for compiling evidence of high quality.

4.2.1. Inclusion criteria

The inclusion criteria encompass the following:

(i) Relevance to AI and COVID-19: Studies were included if they explicitly addressed the deployment of AI technologies in the detection, diagnosis, treatment, or management of COVID-19, or in the analysis of pandemic-related data.
(ii) Peer-reviewed publications: Only peer-reviewed publications were considered, ensuring that all included studies had undergone rigorous academic scrutiny and met the high standards of scientific inquiry.
(iii) Empirical research studies: The review was confined to empirical research studies that presented original data or analyses, providing concrete evidence of AI's efficacy and utility in the pandemic context.

4.2.2. Exclusion criteria

The review employed the exclusion criteria as follows:

(i) Non-English publications: Studies not published in English were excluded, given the linguistic capabilities of the review team and the need to ensure clarity and consistency in the synthesis of findings.
(ii) Preprints and gray literature: Preprints and gray literature were excluded to maintain a focus on validated and peer-reviewed research, thereby upholding the review's standard for evidence-based conclusions.

4.3. Data extraction and analysis

The data extraction and analysis phase are critical in the literature review process, where data is meticulously gathered from selected studies and rigorously analyzed to form meaningful insights. This section elucidates the methodical approach adopted for extracting and analyzing data during the research process.

4.3.1. Data extraction protocol

Data were extracted from studies that met the inclusion criteria, focusing on the application of AI in various aspects of the COVID-19 response globally. This information included data on vaccine efficacy, treatment outcomes, diagnostic accuracy, and predictive analytics. Standardized data extraction forms were employed to ensure consistency and reliability across the data extraction process. These forms were designed to capture all relevant information, including study design, methodology, results, and conclusions.

4.3.2. Analytical framework

The extracted data were synthesized to provide a comprehensive overview of the current state of AI in managing the COVID-19 pandemic. This synthesis involved a qualitative assessment of the findings from the included studies. Where applicable, a quantitative analysis was conducted to ascertain the effectiveness and impact of AI applications. This process involved statistical techniques to combine data from multiple studies, providing a more robust understanding of AI's role in the pandemic.

5. Evolution of AI in healthcare

The evolution of AI in health-care represents a significant shift in medical practice and research. From early rule-based expert systems to deep learning models that leverage vast healthcare data and advanced analytics techniques, AI has found its application in multifaceted areas of healthcare and medicine.[18,19] This section delineates some of the early developments of AI in medical diagnosis, genomics, drug discovery, medical devices, and wearables. These advancements and research have laid a foundation on which current technologies have been honed and adapted in the fight against the COVID-19 pandemic.

5.1. Rule-based expert systems

The inception of AI in health-care can be traced back to the early experiments with rule-based expert systems. One such expert system is MYCIN from the 1970s, designed to diagnose bacterial infections and recommend antibiotics.[20,21] Another significant system was the Internist-I (later developed into CADUCEUS), created in the late 1970s.[22] This system focused on internal medicine and could diagnose complex cases by comparing patient data against a large database of disease profiles. Internist-I's comprehensive approach to diagnosis showcased the potential of AI systems to handle a wide range of medical knowledge. These pioneering efforts established the early relationship between computational algorithms and medical expertise, paving the way for advanced AI applications in modern healthcare, where machine learning and data-driven approaches are now integral.

5.2. Integration of machine learning

The integration of machine learning algorithms marked a significant evolution in AI's application within healthcare. The shift from rule-based systems to data-driven approaches allowed for the analysis of large datasets, leading to more accurate diagnostic tools, personalized treatment plans, and predictive analytics.[23-25] Notably, the development of neural networks and deep learning models has further refined the capabilities of AI, enabling the interpretation of complex medical data with enhanced precision.[26,27]

5.3. AI in genomics and drug discovery

A notable milestone in the evolution of AI in healthcare is its application in genomics.[28-31] and drug discovery.[32,33] The completion of the Human Genome Project in the early 2000s opened new avenues for AI applications in understanding genetic diseases and developing targeted therapies.[34] AI-driven platforms such as AtomNet[35] have since been utilized to identify potential drug candidates, significantly reducing the time and cost associated with traditional drug discovery processes.

5.4. AI-enabled medical devices and wearables

The emergence of AI-enabled medical devices and wearables has significantly benefited patient monitoring and health management. Devices such as smartwatches and fitness trackers, equipped with biomedical sensors and AI algorithms, can now provide real-time insights into an individual's health status, detecting anomalies that may require medical attention.[36] These advancements have not only enhanced preventive healthcare measures but have also empowered individuals to take an active role in managing their health.

5.5. The role of AI in pandemic response

There were no major pandemics before the COVID-19 pandemic where AI was used extensively or prominently in the response. This is primarily because the development and widespread adoption of advanced AI technologies, particularly in healthcare, coincided with or followed the COVID-19 pandemic. Previous health crises, such as the H1N1 influenza pandemic in 2009 or the Ebola outbreak in 2014 – 2016, occurred before AI had reached its current level of sophistication and integration in health-care systems. During these earlier health crises, the use of AI was either very limited or not a significant component of the public health response.

However, it is noteworthy that before COVID-19, research efforts were made to explore the potential use of technology and AI in disease outbreaks.[37] Predictive modeling and data-driven techniques have been studied to predict infectious disease epidemics.[38,39] Other studies demonstrated the use of machine learning analysis of social media and media sources for tracking public health trends and understanding public awareness during health crises.[40,41] These studies collectively illustrate the evolving role of AI, big data, and machine learning in monitoring and predicting disease outbreaks, offering valuable insights for pandemic preparedness and response.

The utilization of AI and big data in managing the COVID-19 pandemic has been unprecedented. The analysis of vast datasets has provided insights that were previously unattainable, demonstrating the evolution of AI and data analytics in the context of pandemics.[42] The COVID-19 pandemic has also been a catalyst for the rapid development and adoption of AI in various aspects of healthcare and public health. This includes areas such as disease detection and diagnosis, vaccine development, treatment strategies, and epidemiological modeling. Significant applications of AI have been identified in the COVID-19 pandemic,[6] building on the results from prior research and the lessons learned from past health crises. The pandemic has highlighted the potential of AI to contribute significantly to managing public health emergencies and is likely to set a precedent for future use in similar scenarios.

6. AI in COVID-19 detection and diagnosis

The COVID-19 pandemic has spurred an unprecedented reliance on AI technologies in disease detection and diagnosis. This section elucidates the multifaceted role of AI in confronting the diagnostic challenges posed by COVID-19, highlighting innovative methodologies and their implications in medical diagnostics.

6.1. Imaging techniques

The integration of AI into imaging techniques has played an important role in the detection and diagnosis of COVID-19.[43] Deep learning models, particularly convolutional neural networks, have been employed to discern patterns in chest X-ray images and computed tomography scans indicative of viral infection.[44,45] Various large datasets of medical images from COVID-19 patients were independently collected for training and validating deep learning models used in detecting COVID-19 in patients.[46,47]

These deep learning models not only detect COVID-19 but also predict and assess the severity of the disease, which is vital for accurate diagnosis and effective patient management. These AI-driven systems can quantify the degree of lung damage, detect signs of pneumonia, and identify other complications associated with severe COVID-19 infections.[48] Advanced imaging techniques have enabled health-care professionals to gauge the extent of lung involvement and other critical factors that classify the severity of the infection.[49] This capability is crucial for triaging patients, determining appropriate levels of care, and making timely decisions regarding treatment strategies.

These AI-driven tools that analyze medical images have demonstrated remarkable efficacy in enhancing the speed and accuracy of COVID-19 diagnosis, thereby alleviating the burden on healthcare systems.

6.2. Machine learning prediction models

A multitude of research studies have investigated the use of machine learning techniques in predicting and detecting COVID-19 based on symptomatology. One notable study in this domain is presented by Ahamad *et al.*,[50] who developed a machine-learning model targeting early-stage symptoms of SARS-CoV-2 infection. Utilizing supervised machine learning methods, they focused on patient characteristics and clinical details such as fever, cough, and lung infection to predict COVID-19 status with over 85% accuracy. Zoabi *et al.*[51] introduced a machine-learning approach using data obtained from tested individuals in Israel. They trained their model on information such as sex, age, exposure to the infected individual, and clinical symptoms recognized at the time of testing. Their model achieved high accuracies in COVID-19 detection and identified key symptoms such as fever and cough as leading indicators for positive diagnosis.

In the paper published by Menni *et al.*,[52] data were obtained from a COVID-19 symptom tracker smartphone app with 2.6 million users in the United States and the United Kingdom. The study found a strong association between the loss of smell and taste and COVID-19-positive cases. Logistic regressions were employed, and a symptom prediction model was developed, showing high sensitivity and specificity in predicting COVID-19.

6.3. NLP in symptom assessment

NLP has been instrumental in the development of AI-based chatbots and virtual health assistants during the COVID-19 pandemic.[53,54] These platforms are capable of conducting preliminary symptom assessments through patient interactions, streamlining the assessment and triage process, and facilitating early detection of potential COVID-19 cases. Interactive digital health assistants, such as Symptoma, have shown to be more accurate than online questionnaires in identifying COVID-19 cases because users can input more detailed information regarding their symptoms through a natural language conversation with the system.[55] By offering accessible and immediate assistance to the public, these tools alleviate the stress and overwhelming volume faced by telephone hotlines and medical institutions.

Furthermore, AI chatbots with advanced NLP capabilities have extended their services to include mental health support. The pandemic has led to increased levels of stress, anxiety, and other mental health issues among the population. Chatbots have provided a first line of psychological support, offering coping strategies, mindfulness exercises, and, in some cases, referral to mental health professionals.[56]

6.4. Wearable Technologies

Wearable technologies have been instrumental in the early detection and symptom monitoring of COVID-19 patients during the pandemic.[57] Wearable devices such as smartwatches and biometric trackers continuously gather physiological and activity data, such as heart rate, daily steps, and sleep patterns. AI systems then analyze this data to detect deviations that may indicate infection, even before clinical symptoms manifest.[58]

AI has emerged as an indispensable tool in the detection and diagnosis of COVID-19. Its application in imaging, symptom assessment, and wearable technology has not only expedited the diagnostic process but also enhanced its precision.

7. AI in COVID-19 treatment and management

The role of AI in the treatment and management of COVID-19, spanning from drug discovery to patient management to telemedicine, has proven instrumental.[59] By leveraging vast datasets, machine learning algorithms, and predictive analytics, AI has enabled healthcare providers to identify potential drugs for treatment, optimize treatment protocols, and improve patient outcomes. The integration of AI in these areas not only enhances the efficiency of healthcare services but also supports the ongoing efforts to control and mitigate the impact of the pandemic. In exploring the various applications of AI in COVID-19 treatment and management, this section highlights the innovative strategies and tools that have been developed and their significant impact on public health responses.

7.1. Drug discovery

AI has played an essential role in expediting the drug discovery process for COVID-19 treatment. Machine learning algorithms have been utilized to predict the structure of the SARS-CoV-2 virus, thereby identifying potential targets for drug therapy.[60,61] Furthermore, AI platforms such as DeepMind's AlphaFold have made significant contributions to understanding the protein folding of the virus, which is crucial for the development of antiviral drugs.[62] The deployment of AI in virtual screening has also allowed researchers to rapidly assess millions of chemical compounds, streamlining the identification of viable drug candidates.[63,64]

7.2. Patient management and monitoring

In the domain of patient management and monitoring, AI systems have been deployed to predict patient outcomes and optimize resource allocation. Predictive analytics have provided healthcare professionals with tools to forecast the progression of the disease in patients, enabling timely interventions.[49] In addition, AI-driven algorithms have been applied to remotely monitor patients' vital signs, thereby reducing the exposure risk for healthcare workers and other patients.[65]

7.3. Telemedicine

Telemedicine, a component of eHealth, involves using information and communication technology to deliver, manage, and monitor health-care services remotely. During the COVID-19 pandemic, telemedicine emerged as a vital tool, especially for patients in isolation.[66] It enabled these patients to receive medical care without risking exposure for themselves or health-care providers to the virus. Furthermore, it alleviated the strain on healthcare facilities, conserved resources such as personal protective equipment, and played a crucial role in the global management of the pandemic.

The surge in demand for healthcare services during the pandemic has underscored the significance of telemedicine, with AI playing a crucial role in its expansion. AI has facilitated remote diagnosis and consultation services, ensuring continuity of care while minimizing the risk of virus transmission.[67] Moreover, AI-powered chatbots have been employed to provide initial medical assessments based on symptoms reported by patients, thus alleviating the strain on medical facilities.[68]

8. AI in COVID-19 prediction and analytics

AI has been utilized in the domain of COVID-19 prediction and analytics as part of the global response to the pandemic. AI models and NLP algorithms have proven pivotal in epidemiological modeling, optimizing resource allocation, and analyzing social media to gauge public sentiment and disseminate information.

8.1. Epidemiological modeling

AI has played a critical role in epidemiological modeling, providing forecasts essential for planning and intervention strategies. Sophisticated machine learning models based on reinforcement learning have been employed to predict the spread of the virus, assess the impact of public health interventions, and estimate the burden on healthcare systems.[69,70] Neural network methods have been implemented to identify COVID-19 clusters, providing insights into how socioeconomic factors and spatial distribution relate to the spread of COVID-19 cases.[71] These models have been crucial in informing government policies, such as implementing lockdowns and organizing vaccination campaigns, to mitigate the spread of the virus.[72]

8.2. Resource allocation

In the realm of resource allocation, AI has been instrumental in ensuring the efficient distribution of medical supplies and medical personnel. Predictive analytics have enabled hospitals to anticipate demand for intensive care units (ICU) and ventilators, facilitating timely procurement and allocation of these critical resources.[73] AI has also been used to develop decision-support tools that assist health-care administrators in making informed decisions about resource distribution, such as determining the need for mechanical ventilation for a COVID-19 patient.[74,75]

8.3. Social media and sentiment analysis

AI has found extensive application on social media platforms for sentiment analysis, misinformation tracking, and understanding public perception regarding COVID-19. NLP algorithms have analyzed vast amounts of data from social media to identify trends in public discourse, monitor compliance with public health measures, and combat the spread of false information.[76] These insights have proven invaluable for public health officials in tailoring communication strategies and effectively addressing public concerns.[77] For example, a study conducted in the United States developed an automatic NLP pipeline to detect potential COVID-19 cases that might have gone untested and unreported, utilizing data generated by Twitter users.[78]

AI has emerged as an indispensable tool in the fight against COVID-19, offering robust solutions for prediction and analytics. The insights gained from AI applications have not only informed public health strategies but have also played a critical role in managing the social dynamics of the pandemic. As we continue to navigate through these challenging times, AI's role in prediction and analytics will evolve and become more deeply integrated into multifarious aspects of pandemic response efforts.

9. Ethical and societal implications

The rapid deployment of AI technologies during the COVID-19 pandemic has given rise to a range of ethical and societal implications that warrant rigorous scrutiny. As AI systems become increasingly integrated into healthcare and public health strategies, concerns surrounding data privacy, algorithmic bias, and accessibility have emerged as critical issues that must be addressed to ensure equitable and ethical technology use.

9.1. Data privacy

The use of AI in managing the COVID-19 pandemic relies heavily on the collection, processing, and analysis of vast amounts of personal data. Contact tracing apps, health monitoring systems, and AI-driven diagnostic tools all operate on inherently personal and sensitive data. It is imperative to protect patient confidentiality and adhere to data protection laws, as breaches can erode public trust and potentially harm individuals.[79] The General Data Protection Regulation (GDPR) in the European Union, along with similar regulations globally, provides a framework for data protection. However, the unprecedented scale of the pandemic poses new challenges in ensuring compliance and safeguarding privacy.[80]

9.2. Algorithmic bias

AI algorithms are susceptible to bias, which can arise from skewed training datasets or flawed design and implementation. In the context of COVID-19, such biases can lead to disparities in diagnosis, treatment, and vaccine distribution, disproportionately affecting marginalized communities.[81] Conducting thorough bias audits and implementing corrective measures are essential to mitigate these risks. The development of AI systems must align with the Findability, Accessibility, Interoperability, and Reusability principles with regard to COVID-19 patient data. In addition, diverse datasets reflecting the heterogeneity of the population should be included.[82]

9.3. Accessibility and inequality

The rapid deployment of AI solutions during the pandemic has highlighted the digital divide and issues of accessibility. Not all populations have equal access to the technologies that facilitate remote healthcare, such as telemedicine, exacerbating existing health inequalities.[83] Furthermore, low-resource settings may lack the infrastructure necessary to implement AI-driven interventions, leading to a disparity in the quality of care and health outcomes.[84] Ensuring equitable access to AI technologies is crucial in the global response to the pandemic and broader healthcare context.[85]

The ethical and societal implications of AI in the COVID-19 era are complex and multifaceted. As we reflect on the challenges posed by the pandemic, it is imperative to foster an ethical AI ecosystem that prioritizes data privacy, mitigates algorithmic bias, and promotes accessibility and equity. Only then can we harness the full potential of AI to serve the greater good without compromising the values of a just and fair society.

10. Case studies

The deployment of AI in response to the COVID-19 pandemic has exhibited significant variation across different countries, resulting in a mix of successes and failures. These case studies provide valuable insights into the potential and limitations of AI in public health emergencies.

10.1. Country-specific implementations

10.1.1. South Korea's AI-powered response

South Korea's response to the COVID-19 pandemic is a prime example of effective AI implementation. The country's swift action in developing AI-driven testing, tracing, and treatment strategies resulted in the efficient containment of the virus. AI algorithms were employed to analyze travel and medical data, facilitating rapid contact tracing and targeted testing.[86-88] Chatbot services such as the Korean COVID-19 chatbot provided citizens with real-time information by integrating public data from the Korea Centers for Disease Control and Prevention and Ministry of Health and Welfare,[89] thereby easing the burden on national health-care hotlines.

10.1.2. Singapore's TraceTogether program

Singapore launched the TraceTogether program, which utilized a mobile application and token-based system to facilitate digital contact tracing.[90] The technology behind the program assessed the proximity and duration of user interactions to notify individuals of potential exposure to the virus. While innovative, the program encountered challenges related to user privacy and data security.[91]

10.1.3. The United States' vaccine distribution

In the United States, AI played a crucial role in optimizing vaccine distribution logistics. Recurrent neural networks helped identify optimal locations for vaccine centers and manage supply chains. However, the reliance on AI also led to some disparities in vaccine allocation, highlighting the need for oversight in AI implementations.[92]

10.2. Success stories and failures

AI-driven diagnostic tools have emerged as a success story, with algorithms such as those developed by DeepMind capable of predicting the structure of proteins associated with SARS-CoV-2, the virus causing COVID-19.[93] This breakthrough holds implications for understanding the virus's mechanisms and developing treatments.

Moreover, AI has proven successful in disseminating public health messaging via social media platforms, chatbots, and other digital means. These AI systems have effectively tailored messages to specific demographics, thereby improving public engagement and compliance with health guidelines.[94]

Conversely, some AI predictive models have failed to provide accurate forecasts for the spread of the virus. In many instances, these models were unable to account for the dynamic nature of human behavior and policy changes, leading to over- or under-estimation of case numbers.[95]

These case studies underscore the importance of careful management of AI applications in pandemic response efforts. Success depends not only on the technology itself but also on factors such as data quality, user engagement, and the ethical use of AI.

11. Future directions

The COVID-19 pandemic has accelerated the integration of AI in healthcare and public health. Looking ahead, several emerging technologies and policy recommendations could shape the next phase of AI in pandemic preparedness and response.

11.1. Emerging technologies

Quantum computing holds the promise of processing complex datasets much faster than traditional computers. In the context of pandemics, quantum algorithms could revolutionize the way we model viral spread, optimize supply chains for medical supplies, and discover new therapeutic drugs.[96]

Next-generation sequencing (NGS) technologies are rapidly evolving, allowing for quicker and more affordable genomic sequencing. AI, combined with NGS, could enable real-time tracking of pathogen evolution, helping public health officials stay ahead of mutations and variants of concern.[97]

Blockchain technology offers a secure and transparent way to manage health data. In pandemics, blockchain can ensure the integrity of health records, facilitate secure data sharing for AI algorithms, and support contact tracing efforts without compromising privacy.[98]

11.2. Policy recommendations

Robust data governance frameworks are essential to ensure that AI systems have access to high-quality, representative data while safeguarding individual privacy. Policies must be developed to address data ownership, consent, and anonymization.[99]

Given the global nature of pandemics, international cooperation is imperative. Policy recommendations should encourage the sharing of AI technologies and expertise across borders, as well as fostering collaborative efforts in research and development.[100]

To fully harness the potential of AI, investments in education and workforce development are crucial. This effort includes training healthcare professionals in AI applications and promoting AI literacy among the general population.[101]

The future of AI in the context of pandemics is promising, with emerging technologies offering new tools to combat infectious diseases. However, realizing this potential will require thoughtful policy recommendations that promote innovation while addressing ethical, legal, and social implications.

12. Conclusion

The COVID-19 pandemic has served as a catalyst for unprecedented global change, particularly in the realms of healthcare and technology. AI has emerged as a critical tool in combating the pandemic, offering solutions for detection, diagnosis, treatment, and management of the disease. In addition, it has played a significant role in understanding and predicting the spread of the virus, aiding in resource allocation, and analyzing public sentiment.

Reflecting on the lessons learned, it becomes evident that AI holds the potential to transform public health responses to future pandemics. However, this potential can only be realized through ethical practices, equitable access, and international collaboration. The integration of AI in healthcare demands a commitment to data privacy, a focus on reducing algorithmic bias, and an emphasis on the creation of systems accessible to all, regardless of socioeconomic status.

The case studies presented throughout this review highlight both the successes and failures of AI implementations in various contexts, offering valuable insights for future endeavors. Moving forward, emerging technologies such as quantum computing, blockchain, and NGS will further enhance the capabilities of AI in public health.

Acknowledgments

None.

Funding

This research was funded by Research on Quality Assurance and Evaluation of Higher Education in Jiangsu Province under Grant No. 2023JSETKT032.

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Emma Yann Zhang, Adrian David Cheok
Formal analysis: Emma Yann Zhang, Adrian David Cheok
Investigation: Emma Yann Zhang, Adrian David Cheok
Methodology: Emma Yann Zhang, Adrian David Cheok
Writing – original draft: Emma Yann Zhang, Adrian David Cheok
Writing – review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable.

References

1. Marques ES, Moraes CL, Hasselmann MH, Deslandes SF, Reichenheim ME. Violence against women, children, and adolescents during the COVID-19 pandemic: Overview, contributing factors, and mitigating measures. *Cad Saude Publica*. 2020;36(4):e00074420.

   doi: 10.1590/0102-311X00074420

2. Maital S, Barzani E. The global economic impact of COVID-19: A summary of research. *Samuel Neaman Inst Nat Policy Res*. 2020;2020:1-12.

3. Khan KS, Mamun MA, Griffiths MD, Ullah I. The mental health impact of the COVID-19 pandemic across different cohorts. *Int J Ment Health Addict*. 2022;20:380-386.

   doi: 10.1007/s11469-020-00367-0

4. Msemburi W, Karlinsky A, Knutson V, Aleshin-Guendel S, Chatterji S, Wakefield J. The WHO estimates of excess mortality associated with the COVID-19 pandemic. *Nature*. 2023;613:130-137.

   doi: 10.1038/s41586-022-05522-2

5. Assefa Y, Gilks CF, Pas R, Reid S, Gete DG, Van Damme W. Reimagining global health systems for the 21st century: lessons from the COVID-19 pandemic. *BMJ Glob Health*. 2021;6:e004882.

   doi: 10.1136/bmjgh-2021-004882

6. Vaishya R, Javaid M, Khan IH, Haleem A. Artificial intelligence (AI) applications for COVID-19 pandemic. *Diabetes Metab Syndr Clin Res Rev*. 2020;14:337-339.

   doi: 10.1016/j.dsx.2020.04.012

7. Darapaneni N, Sreevanth AT, Paduri AR, *et al.* Explainable Diagnosis, Lesion Segmentation and Quantification of COVID-19 Infection from CT Images using Convolutional Neural Networks. In: *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE; 2022. p. 171-178.

   doi: 10.1109/IEMCON53756.2022.9623045

8. Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in the diagnosis of COVID-19: Challenges and perspectives. *Int J Biol Sci*. 2021;17:1581.

   doi: 10.7150/ijbs.58855

9. Silva C, Saraee D. Literature review on epidemiological modelling, spatial modelling and artificial intelligence for COVID-19. *J Adv Med Med Res*. 2021;33:8-21.

   doi: 10.9734/jammr/2021/v33i530841

10. Vicari R, Komendatova N. Systematic meta-analysis of research on AI tools to deal with misinformation on social media during natural and anthropogenic hazards and disasters. *Humanit Soc Sci Commun*. 2023;10:1-14.

    doi: 10.1057/s41599-023-01838-0

11. Al Sulais E, Mosli M, AlAmeel T. The psychological impact of COVID-19 pandemic on physicians in Saudi Arabia: A cross-sectional study. *Saudi J Gastroenterol*. 2020;26:249.

    doi: 10.4103/sjg.SJG_173_20

12. Poon YS, Lin YP, Griffiths P, Yong KK, Seah B, Liaw SY. A global overview of healthcare workers' turnover intention amid COVID-19 pandemic: A systematic review with future directions. *Hum Resour Health*. 2022;20:70.

    doi: 10.1186/s12960-022-00764-7

13. Mhlanga D. The role of artificial intelligence and machine learning amid the COVID-19 pandemic: What lessons are we learning on 4IR and the sustainable development goals. *Int J Environ Res Public Health*. 2022;19:1879.

    doi: 10.3390/ijerph19041879

14. Yogi MK, Garikipati J. Future scope of artificial intelligence in healthcare for COVID-19. In: *Emerging Technologies for Combatting Pandemics*. United Kingdom: Taylor & Francis; 2022. p. 85-100.

    doi: 10.1201/9781003324447-5

15. Pham QV, Nguyen DC, Huynh-The T, Hwang WJ, Pathirana PN. Artificial intelligence (AI) and big data for coronavirus (COVID-19) pandemic: A survey on the state-of-the-arts. *IEEE Access*. 2020;8:130820-130839.

    doi: 10.1109/ACCESS.2020.3009328

16. Naudé W. *Artificial Intelligence against COVID-19: An Early Review. IZA Discussion Paper No. 13110*; 2020.

    doi: 10.2139/ssrn.3568314

17. Adly AS, Adly AS, Adly MS. Approaches based on artificial intelligence and the internet of intelligent things to prevent the spread of COVID-19: Scoping review. *J Med Internet Res*. 2020;22:e19104.

    doi: 10.2196/19104

18. Jiang F, Jiang Y, Zhi H, *et al*. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc Neurol*. 2017;2:230-243.

    doi: 10.1136/svn-2017-000101

19. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism*. 2017;69:S36-S40.

    doi: 10.1016/j.metabol.2017.01.011

20. Shortliffe EH. *MYCIN: A Rule-Based Computer Program for Advising Physicians Regarding Antimicrobial Therapy Selection. PhD Thesis. Stanford University*; 1974.

    doi: 10.1145/1408800.1408906

21. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Comput Biomed Res*. 1975;8:303-320.

    doi: 10.1016/0010-4809(75)90009-9

22. Miller RA, Pople HE Jr., Myers JD. Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. In: *Computer-assisted Medical Decision Making*. Berlin: Springer; 1985. p. 139-158.

    doi: 10.1007/978-1-4612-5108-8_8

23. Bradley AP. *Machine Learning for Medical Diagnostics: Techniques for Feature Extraction, Classification, and Evaluation*. Australia: The University of Queensland; 1996.

24. Kononenko I. Inductive and Bayesian learning in medical diagnosis. *Appl Artif Intell*. 1993;7:317-337.

    doi: 10.1080/08839519308949977

25. Zupan B, Demšar J, Kattan MW, Beck JR, Bratko I. Machine learning for survival analysis: A case study on recurrence of prostate cancer. *Artif Intell Med*. 2000;20:59-75.

    doi: 10.1016/S0933-3657(00)00053-1

26. Miller AS, Blott BH, Hames TK. Review of neural network applications in medical imaging and signal processing. *Med Biol Eng C*omput. 1992;30:449-464.

    doi: 10.1007/BF02441652

27. Lo SCB, Chan HP, Lin JS, Li H, Freedman MT, Mun SK. Artificial convolution neural network for medical image pattern recognition. *Neural Netw*. 1995;8:1201-1214.

    doi: 10.1016/0893-6080(95)00061-5

28. Larranaga P, Calvo B, Santana R, *et al*. Machine learning in bioinformatics. *Brief Bioinform*. 2006;7:86-112.

    doi: 10.1093/bib/bbk007

29. Dubitzky W, Granzow M, Berrar DP. *Fundamentals of Data Mining in Genomics and Proteomics*. Berlin: Springer Science & Business Media; 2007.

    doi: 10.1007/978-0-387-47509-7

30. Hayes WS, Borodovsky M. How to interpret an anonymous bacterial genome: Machine learning approach to gene identification. *Genome Res*. 1998;8:1154-1171.

    doi: 10.1101/gr.8.11.1154

31. Zhavoronkov A, Ivanenkov YA, Aliper A, *et al*. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol*. 2019;37:1038-1040.

    doi: 10.1038/s41587-019-0224-x

32. Burbidge R, Trotter M, Buxton B, Holden S. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput Chem*. 2001;26:5-14.

doi: 10.1016/S0097-8485(01)00094-8

33. Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J Chem Inf Comput Sci*. 2003;43:2048-2056.

doi: 10.1021/ci0341161

34. Collins FS, Morgan M, Patrinos A. The human genome project: Lessons from large-scale biology. *Science*. 2003;300:286-290.

doi: 10.1126/science.1084564

35. Wallach I, Dzamba M, Heifets A. AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv:1510.02855. 2015.

36. Piwek L, Ellis DA, Andrews S, Joinson A. The rise of consumer health wearables: Promises and barriers. *PLoS Med*. 2016;13:e1001953.

doi: 10.1371/journal.pmed.1001953

37. Kaur G. Pandemic management via technology: A review. *Management*. 2011;40:181-187.

doi: 10.1016/j.indmarman.2010.06.026

38. Sadilek A, Kautz H, Silenzio V. Predicting Disease Transmission from Geo-Tagged Micro-Blog Data. Vol. 26. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2012. p. 136-142.

doi: 10.1609/aaai.v26i1.8103

39. Raza Abidi SS, Goh A. Applying Knowledge Discovery to Predict Infectious Disease Epidemics. In: *Pacific Rim International Conference on Artificial Intelligence*. Berlin: Springer; 1998. p. 170-181.

doi: 10.1007/BFb0095267

40. Lampos V, Cristianini N. Tracking the Flu Pandemic by Monitoring the Social Web. In: *2010 2nd International Workshop on Cognitive Information Processing*. IEEE; 2010. p. 411-416.

doi: 10.1109/CIP.2010.5604088

41. Choi S, Lee J, Kang MG, Min H, Chang YS, Yoon S. Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks. *Methods*. 2017;129:50-59.

doi: 10.1016/j.ymeth.2017.04.004

42. Bragazzi NL, Dai H, Damiani G, Behzadifar M, Martini M, Wu J. How big data and artificial intelligence can help better manage the COVID-19 pandemic. *Int J Environ Res Public Health*. 2020;17:3176.

doi: 10.3390/ijerph17093176

43. Dong D, Tang Z, Wang S, *et al*. The role of imaging in the detection and management of COVID-19: A review. *IEEE Rev Biomed Eng*. 2020;14:16-29.

doi: 10.1109/RBME.2020.2990959

44. Wang L, Lin ZQ, Wong A. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Sci Rep.* 2020;10:19549.

doi: 10.1038/s41598-020-76550-z

45. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya UR. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput Biol Med*. 2020;121:103792.

doi: 10.1016/j.compbiomed.2020.103792

46. Li L, Qin L, Xu Z, *et al*. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology*. 2020;296:E65-E71.

doi: 10.1148/radiol.2020200905

47. Shi F, Xia L, Shan F, *et al*. Large-scale screening to distinguish between COVID-19 and community-acquired pneumonia using infection size-aware classification. *Phys Med Biol*. 2021;66:065031.

doi: 10.1088/1361-6560/abe838

48. Zhang K, Liu X, Shen J, *et al*. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell*. 2020;181:1423-1433.e11.

doi: 10.1016/j.cell.2020.04.045

49. Jiang X, Coffee M, Bari A, *et al*. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Comput Mater Continua*. 2020;63:537-551.

doi: 10.32604/cmc.2020.010691

50. Ahamad MM, Aktar S, Rashed-Al-Mahfuz M, *et al*. A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. *Expert Syst Appl*. 2020;160:113661.

doi: 10.1016/j.eswa.2020.113661

51. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digit Med*. 2021;4:3.

doi: 10.1038/s41746-020-00372-6

52. Menni C, Valdes AM, Freydin MB, *et al*. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat Med*. 2020;26:1037-1040.

doi: 10.1038/s41591-020-0916-2

53. Miner AS, Laranjo L, Kocaballi AB. Chatbots in the fight against the COVID-19 pandemic. *NPJ Digit Med*. 2020;3:65.

    doi: 10.1038/s41746-020-0280-0

54. Almalki M, Azeez F. Health chatbots for fighting COVID-19: A scoping review. *Acta Inform Med*. 2020;28:241.

    doi: 10.5455/aim.2020.28.241-247

55. Martin A, Nateqi J, Gruarin S, *et al.* An artificial intelligence-based first-line defence against COVID-19: Digitally screening citizens for risks via a chatbot. *Sci Rep.* 2020;10:19012.

    doi: 10.1038/s41598-020-75912-x

56. Boucher EM, Harake NR, Ward HE, *et al*. Artificially intelligent chatbots in digital mental health interventions: A review. *Expert Rev Med Devices*. 2021;18:37-49.

    doi: 10.1080/17434440.2021.2013200

57. Channa A, Popescu N, Skibinska J, Burget R. The rise of wearable devices during the COVID-19 pandemic: A systematic review. *Sensors*. 2021;21:5787.

    doi: 10.3390/s21175787

58. Mishra T, Wang M, Metwally AA, *et al*. Pre-symptomatic detection of COVID-19 from smartwatch data. *Nat Biomed Eng*. 2020;4:1208-1220.

    doi: 10.1038/s41551-020-00640-6

59. Gandla K, Reddy KTK, Babu PV, Sagapola R, Sudhakar P. A review of artificial intelligence in treatment of COVID-19. *J Pharm Negat Results*. 2022;13:254-264.

60. Smith MD, Smith JC. Repurposing therapeutics for COVID-19: Supercomputer-Based Docking to the SARS-CoV-2 Viral Spike Protein and Viral Spike Protein-Human ACE2 Interface. ChemRxiv. 2020.

    doi: 10.26434/chemrxiv.11871402

61. Zhou Y, Wang F, Tang J, Nussinov R, Cheng F. Artificial intelligence in COVID-19 drug repurposing. *Lancet Digit Health*. 2020;2:e667-e676.

    doi: 10.1016/S2589-7500(20)30192-8

62. Senior AW, Evans R, Jumper J, *et al*. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577:706-710.

    doi: 10.1038/s41586-019-1923-7

63. Jang WD, Jeon S, Kim S, Lee SY. Drugs repurposed for COVID-19 by virtual screening of 6,218 drugs and cell-based assay. *Proc Natl Acad Sci U S A*. 2021;118:e2024302118.

    doi: 10.1073/pnas.2024302118

64. Kandeel M, Al-Nazawi M. Virtual screening and repurposing of FDA approved drugs against COVID-19 main protease. *Life Sci*. 2020;251:117627.

    doi: 10.1016/j.lfs.2020.117627

65. Rohmetra H, Raghunath N, Narang P, Chamola V, Guizani M, Lakkaniga NR. AI-enabled remote monitoring of vital signs for COVID-19: methods, prospects and challenges. *Computing*. 2021;105:1-27.

    doi: 10.1007/s00607-021-00937-7

66. Bokolo AJ. Application of telemedicine and eHealth technology for clinical services in response to COVID-19 pandemic. *Health Technol* (*Berl*). 2021;11:359-366.

    doi: 10.1007/s12553-020-00516-4

67. Webster P. Virtual health care in the era of COVID-19. *Lancet*. 2020;395:1180-1181.

    doi: 10.1016/S0140-6736(20)30818-7

68. Judson TJ, Odisho AY, Young JJ, *et al*. Implementation of a digital chatbot to screen health system employees during the COVID-19 pandemic. *J Am Med Inform Assoc*. 2020;27:1450-1455.

    doi: 10.1093/jamia/ocaa130

69. Chadi MA, Mousannif H. A reinforcement learning based decision support tool for epidemic control: Validation study for COVID-19. *Appl Artif Intell*. 2022;36:2031821.

    doi: 10.1080/08839514.2022.2031821

70. Guo X, Chen P, Liang S, *et al*. PaCAR: COVID-19 pandemic control decision making via large-scale agent-based modeling and deep reinforcement learning. *Med Decis Making*. 2022;42:1064-1077.

    doi: 10.1177/0272989X221107902

71. Ghahramani M, Pilla F. Leveraging artificial intelligence to analyze the COVID-19 distribution pattern based on socioeconomic determinants. *Sustain Cities Soc*. 2021;69:102848.

    doi: 10.1016/j.scs.2021.102848

72. Narayan K, Rathore H, Znidi F. Using epidemic modeling, machine learning and control feedback strategy for policy management of COVID-19. *IEEE Access*. 2022;10:98244-98258.

    doi: 10.1109/ACCESS.2022.3206790

73. Wang H. The application of artificial intelligence in health care resource allocation before and during the COVID-19 pandemic: Scoping Health Policy. *JMIR AI*. 2023;5:6.

74. Yu L, Halalau A, Dalal B, *et al*. Machine learning methods to predict mechanical ventilation and mortality in patients with COVID-19. *PLoS One*. 2021;16:e0249285.

    doi: 10.1371/journal.pone.0249285

75. Douville NJ, Douville CB, Mentz G, *et al*. Clinically applicable approach for predicting mechanical ventilation in patients with COVID-19. *Br J Anaesth*. 2021;126:578-589.

    doi: 10.1016/j.bja.2020.11.034

76. Cinelli M, Quattrociocchi W, Galeazzi A, *et al*. The COVID-19 social media infodemic. *Sci Rep*. 2020;10:16598.

doi: 10.1038/s41598-020-73510-5

77. Pulido CM, Villarejo-Carballido B, Redondo-Sama G, Gómez A. COVID-19 infodemic: More retweets for science-based information on coronavirus than for false information. *Int Sociol*. 2020;35:377-392.

doi: 10.1177/0268580920914755

78. Klein AZ, Magge A, O'Connor K, Flores AIJ, Weissenbacher D, Gonzalez Hernandez G. Toward using Twitter for tracking COVID-19: A natural language processing pipeline and exploratory data set. *J Med Internet Res*. 2021;23:e25314.

doi: 10.2196/25314

79. Newlands G, Lutz C, Tamò-Larrieux A, Villaronga EF, Harasgama R, Scheitlin G. Innovation under pressure: Implications for data privacy during the Covid-19 pandemic. *Big Data Soc*. 2020;7:2053951720976680.

doi: 10.1177/2053951720976680

80. Christofidou M, Lea N, Coorevits P. A literature review on the GDPR, COVID-19 and the ethical considerations of data protection during a time of crisis. *Yearb Med Inform*. 2021;30:226-232.

doi: 10.1055/s-0041-1726512

81. Delgado J, Manuel A, Parra I, *et al*. Bias in algorithms of AI systems developed for COVID-19: A scoping review. *J Bioeth Inq*. 2022;19:407-419.

doi: 10.1007/s11673-022-10200-z

82. Queralt-Rosinach N, Kaliyaperumal R, Bernabé CH, *et al*. Applying the FAIR principles to data in a hospital: Challenges and opportunities in a pandemic. *J Biomed Semant*. 2022;13:12.

doi: 10.1186/s13326-022-00263-7

83. Laurencin CT, McClinton A. The COVID-19 pandemic: A call to action to identify and address racial and ethnic disparities. *J Racial Ethn Health Disparities*. 2020;7:398-402.

doi: 10.1007/s40615-020-00756-0

84. Garcia Elorrio E, Arrieta J, Arce H, *et al*. The COVID-19 pandemic: A call to action for health systems in Latin America to strengthen quality of care. *Int J Qual Health Care*. 2021;33:mzaa062.

doi: 10.1093/intqhc/mzaa062

85. Manjarrés Á, Fernández-Aller C, López-Sánchez M, Rodríguez-Aguilar JA, Sierra Castañer M. Artificial intelligence for a fair, just, and equitable world. *IEEE Technol Soc Mag*. 2021;40:19-24.

doi: 10.1109/MTS.2021.3056292

86. Sinha A, Rathi M. COVID-19 prediction using AI analytics for South Korea. *Appl Intell* (*Dordr*). 2021;51:8579-8597.

doi: 10.1007/s10489-021-02352-z

87. Heo K, Lee D, Seo Y, Choi H. Searching for digital technologies in containment and mitigation strategies: Experience from South Korea COVID-19. *Ann Glob Health*. 2020;86:109.

doi: 10.5334/aogh.2993

88. Chung H, Ko H, Kang WS, *et al*. Prediction and feature importance analysis for severity of COVID-19 in South Korea using artificial intelligence: Model development and validation. *J Med Internet Res*. 2021;23:e27060.

doi: 10.2196/27060

89. Nam T. How did Korea use technologies to manage the COVID-19 crisis? A country report. *Int Rev Public Admin*. 2020;25:225-242.

doi: 10.1080/12294659.2020.1848061

90. Lee JK, Lin L, Kang H. The influence of normative perceptions on the uptake of the COVID-19 TraceTogether digital contact tracing system: Cross-sectional study. *JMIR Public Health Surveill*. 2021;7:e30462.

doi: 10.2196/30462

91. Stevens H, Haines MB. Tracetogether: Pandemic response, democracy, and technology. *East Asian Sci Technol Soc*. 2020;14:523-532.

doi: 10.1215/18752160-8698301

92. Davahli MR, Karwowski W, Fiok K. Optimizing COVID-19 vaccine distribution across the United States using deterministic and stochastic recurrent neural networks. *PLoS One*. 2021;16:e0253925.

doi: 10.1371/journal.pone.0253925

93. Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583-589.

doi: 10.1038/s41586-021-03819-2

94. Gunasekeran DV, Tseng RMWW, Tham YC, Wong TY. Applications of digital health for public health responses to COVID-19: A systematic scoping review of artificial intelligence, telehealth and related technologies. *NPJ Digital Med*. 2021;4:40.

doi: 10.1038/s41746-021-00412-9

95. Ioannidis JPA, Cripps S, Tanner MA. Forecasting for COVID-19 has failed. *Int J Forecast*. 2022;38:423-438.

doi: 10.1016/j.ijforecast.2020.08.004

96. Jayanthi P, Rai BK, Muralikrishna I. The potential of quantum computing in healthcare. In: *Technology Road Mapping for Quantum Computing and Engineering*. United States: IGI Global; 2022. p. 81-101.

doi: 10.4018/978-1-7998-9183-3.ch006

97. John G, Sahajpal NS, Mondal AK, *et al*. Next-generation sequencing (NGS) in COVID-19: A tool for SARS-CoV-2 diagnosis, monitoring new strains and phylodynamic modeling in molecular epidemiology. *Curr Issues Mol Biol*.

2021;43:845-867.

doi: 10.3390/cimb43020061

98. Ng WY, Tan TE, Movva PVH, *et al*. Blockchain applications in health care for COVID-19 and beyond: A systematic review. *Lancet Digit Health*. 2021;3:e819-e829.

doi: 10.1016/S2589-7500(21)00210-7

99. Juddoo S, George C, Duquenoy P, Windridge D. Data governance in the health industry: Investigating data quality dimensions within a big data context. *Appl Syst Innov.* 2018;1:43.

doi: 10.3390/asi1040043

100. Bernardo T, Sobkowich KE, Forrest RO, *et al*. Collaborating in the time of COVID-19: The scope and scale of innovative responses to a global pandemic. *JMIR Public Health Surveill*. 2021;7:e25935.

doi: 10.2196/25935

101. Dunn P, Hazzard E. Technology approaches to digital health literacy. *Int J Cardiol*. 2019;293:294-296.

doi: 10.1016/j.ijcard.2019.06.039

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties

Ummara Mumtaz[1], Awais Ahmed[2], and Summaya Mumtaz[1]*

[1]Department of Information Technology, University of the Cumberlands, Williamsburg, Kentucky, United States of America

[2]Department of Gynecology and Obstetrics, University of Concepción, Concepción, Chile

Abstract

The purpose of this review is to provide a comprehensive overview of the latest advancements in utilizing large language models (LLMs) in the health-care sector, emphasizing their transformative impact across various medical domains. LLMs have become pivotal in supporting healthcare, including physicians, health-care providers, and patients. Our review provides insight into the applications of LLMs in healthcare, specifically focusing on diagnostic and treatment-related functionalities. We shed light on how LLMs are applied in cancer care, dermatology, dental care, neurodegenerative disorders, and mental health, highlighting their innovative contributions to medical diagnostics and patient care. Throughout our analysis, we explore the challenges and opportunities associated with integrating LLMs in healthcare, recognizing their potential across various medical specialties despite existing limitations. In addition, we offer an overview of handling diverse data types within the medical field.

Keywords: Large language models; Medical specialties; Cancer; Mental health; Healthcare; Diagnosis and treatments; Clinical notes; Dermatology

***Corresponding author:**
Summaya Mumtaz
(summaya.mumtaz@gmail.com)

Citation: Mumtaz U, Ahmed A, Mumtaz S. LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties. *Artif Intell Health*. 2024;1(2): 16-28. doi: 10.36922/aih.2558

Received: December 28, 2023

Accepted: February 23, 2024

Published Online: April 2, 2024

1. Introduction

The field of artificial intelligence (AI) has undergone a remarkable evolution in recent years, with significant advancements, particularly noticeable in natural language processing (NLP) and the development of large language models (LLMs). These models represent a paradigm shift in AI's capability to understand, generate, and interact using human language. At their foundation, LLMs are complex algorithms trained on vast, text-based documents and datasets.[1] Such extensive training allows them to recognize patterns adeptly, predict subsequent words in a sentence, and generate coherent, contextually relevant text for the specified inputs, often called prompts within the NLP community. This ability demonstrates the technical prowess of LLMs and signifies their potential to revolutionize how machines understand and process human language. One of the most prominent features of LLMs is their proficiency in processing and analyzing large volumes of text rapidly and accurately, a capability that far surpasses human potential in speed and efficiency.[2] This quality makes them indispensable in areas requiring the analysis of extensive data sets. They are also known as "few-shot" learners, meaning once trained on massive datasets, they can be retrained for new domains utilizing a small number of domain-specific examples.[3]

LLMs have become increasingly prevalent in the medical domain, due to their versatility, and expanding influence. Their applications in healthcare are multifaceted, ranging from processing vast quantities of medical data and interpreting clinical notes to generating comprehensive, human-readable reports.[4] This broad spectrum of functionalities shows how LLMs are not just tools for data processing but are also instrumental in providing innovative solutions across various aspects of healthcare. LLMs are increasingly being utilized to tackle critical challenges in patient care. This includes providing customized educational content to patients, assisting health-care professionals in making complex diagnostic decisions, and easing the administrative burdens often associated with health-care provision.[4,5]

While LLMs have been applied across a spectrum of activities in healthcare, including medical question answering, examination, pure research-oriented tasks, and administrative duties in hospitals, this review will focus exclusively on their practical applications in healthcare, such as diagnostics and treatment purposes. We uncover their deployment in critical areas such as cancer care, dermatology, dental, mental health, and other core medical specialties listed in Figure 1. This exploration is crucial, as it showcases LLMs' capacity to innovate and streamline medical diagnostics, patient care, treatment tasks, and also address the challenges and opportunities in harnessing their full potential in complex medical areas. In this review, we conduct an in-depth analysis of the applications of LLMs across different medical fields. We focus on the advancements and challenges of integrating these sophisticated models into routine health-care practices. We offer insights into the current state of progress and identify barriers to their widespread adoption in clinical settings. The paper is structured to cover each medical specialty and associated challenges, followed by examining various data types in the medical field. The conclusion summarizes the findings and implications.

2. Cancer care (oncology)

Cancer is characterized by the uncontrolled growth of abnormal cells in the body, a topic encompassed under the big umbrella discipline called oncology – the study of cancer types and related factors. Adopting LLMs such as ChatGPT in oncology has become a focal point of recent research, especially in supporting decision-making processes for cancer treatment. These advanced models are being explored for their capability to enhance diagnostic accuracy, personalize therapy options, and streamline patient care in oncology. By analyzing vast amounts of data, LLMs can provide insights that potentially improve treatment outcomes and patient management strategies. In the subsequent discussion, we explore the studies dedicated to integrating LLMs within oncological care, encapsulating the innovative efforts to harness LLMs' capabilities in enhancing the diagnostic, treatment, and management processes associated with cancer care.

In a study conducted by Sorin *et al.*,[6] the capabilities of ChatGPT, an LLM were explored as a decision-support tool for breast tumor boards. The research's primary objective was determining how ChatGPT's recommendations align with expert-driven decisions during breast tumor board meetings. For this purpose, clinical data from ten patients discussed in a breast tumor board at their institution were inputted into ChatGPT-3.5. Subsequently, the model's management recommendations were compared with the final decisions made by the tumor board. Moreover, two senior radiologists independently evaluated ChatGPT's responses, grading them on a scale from 1 (complete disagreement) to 5 (complete agreement) across three categories: summarization of the case, the recommendation provided, and the explanation for that recommendation. Most patients in the study (80%) had invasive ductal carcinoma, with one case each of ductal carcinoma *in situ* and a phyllodes tumor with atypia. ChatGPT's recommendations aligned with the tumor board's decisions in seven out of the ten cases, marking a 70% concordance. On grading, the first reviewer gave mean scores of 3.7, 4.3, and 4.6 for summarization, recommendation, and explanation, respectively, while the second reviewer's scores were 4.3, 4.0, and 4.3 in the same categories. As an initial exploration, the study suggests that LLMs like ChatGPT are potentially valuable tools for breast tumor boards. However, as technology rapidly advances, medical professionals must know its advantages and potential limitations.

In a study by Lukac *et al.*[7] in January 2023, the capabilities of ChatGPT to assist in the decision-making process for therapy planning in primary breast cancer cases were investigated. Although the ChatGPT was able to identify specific risk factors for hereditary breast cancer and could discern elderly patients requiring chemotherapy assessment for cost/benefit evaluation, it generally offered non-specific recommendations concerning various treatment modalities such as chemotherapy and radiation therapy. Notably, it made errors in patient-specific therapy suggestions, misidentifying patients with Her2 1+ and 2+ (FISH negative) as candidates for trastuzumab therapy and mislabeling endocrine therapy as "hormonal treatment." The study concluded that while ChatGPT demonstrates potential utility in clinical medicine, its current version lacks the precision to offer specific therapy recommendations for primary breast cancer patients. It underscores the necessity for further refinement before it can be a reliable adjunct in multidisciplinary tumor board decisions.

Gebrael *et al.*[8] assessed the utility of ChatGPT 4.0 to enhance triage efficiency and accuracy in emergency rooms for patients with metastatic prostate cancer. Between May 2022 and April 2023, clinical data of 147 patients presenting with metastatic prostate cancer were examined, of which 56 were selected based on inclusion criteria. ChatGPT demonstrated a high sensitivity of 95.7% for determining patient admissions but had a low specificity of 18.2% for discharges. It agreed with physicians' primary diagnoses in 87.5% of cases. It outperformed physicians regarding accurate terminology usage (42.9% vs. 21.4%) and diagnosis comprehensiveness, having a median diagnosis count of 3 compared to physicians' 2. ChatGPT was more concise in its responses and provided more additional treatment recommendations than physicians. The data suggest that the ChatGPT could serve as a valuable tool for assisting medical professionals in emergency room settings, potentially enhancing triage efficiency and the overall quality of patient care.

A study led Rao *et al.*[9] investigated the potential of ChatGPT-3.5 and GPT-4 (OpenAI) in aiding radiologic decision-making, specifically focusing on breast cancer screening and breast pain imaging services. The researchers measured the models' responses against the ACR Appropriateness Criteria using two prompt formats: "open-ended" (OE) and "select all that apply" (SATA). For breast cancer screening, both versions scored an average of 1.830 (out of 2) in the OE format, but GPT-4 outperformed ChatGPT-3.5 in the SATA format, achieving 98.4% accuracy compared to 88.9%. Regarding breast pain, GPT-4 again showed superiority, registering an average OE score of 1.666 and 77.7% in SATA, while ChatGPT-3.5 scored 1.125% and 58.3%, respectively. The data suggest the growing viability of LLMs like ChatGPT in enhancing radiologic decision-making processes, with potential benefits for clinical workflows and more efficient radiological services. However, further refinement and broader application cases are needed for full validation.

Hana *et al.*[10] conducted a retrospective study to evaluate the appropriateness of ChatGPT's responses to common questions concerning breast cancer prevention and screening. By leveraging methodologies from prior research that assessed ChatGPT's capacity to address cardiovascular disease-related inquiries, the team formulated 25 questions rooted in the BI-RADS Atlas and their clinical experiences within tertiary care breast imaging departments. Each question was posed to ChatGPT three times, and three fellowship-trained breast radiologists critically assessed the responses. The radiologists categorized each response as "appropriate," "inappropriate," or "unreliable" based on the content's clinical relevance and consistency. Their evaluations considered two hypothetical scenarios: content for a hospital website and direct chatbot-patient interactions. The majority's opinion dictated the final determination of appropriateness. Their results revealed that ChatGPT provided suitable answers for 88% (22 out of 25) of the questions in both contexts. However, one question pertained to mammography scheduling in light of COVID-19 vaccination, which elicited an inappropriate response.

In addition, there were inconsistencies in answers related to breast cancer prevention and screening location queries. While ChatGPT frequently referenced guidelines from the American Cancer Society in its responses, it omitted those from the American College of Radiology and the U. S. Preventive Services Task Force. These findings aligned with earlier research by Sarraju *et al.*,[11] where 84% of ChatGPT's cardiovascular disease prevention responses were deemed appropriate. Despite considerable potential as an automated tool for patient education on breast cancer, ChatGPT exhibited certain limitations, emphasizing the essential role of physician oversight and the ongoing need for further refinement and research into LLMs in health-care education.

Schulte,[12] in 2023, explored the ability of ChatGPT to identify suitable treatments for advanced solid cancers. Through a structured approach, the study assessed ChatGPT's capacity to list appropriate systemic therapies for newly diagnosed advanced solid malignancies and then compared the treatments ChatGPT suggested with those recommended by the National Comprehensive Cancer Network (NCCN) guidelines. This comparison resulted in the valid therapy quotient (VTQ) measure. The research encompassed 51 diagnoses and found that ChatGPT could identify 91 unique medications related to advanced solid tumors. On average, the VTQ was 0.77, suggesting a reasonably high agreement between ChatGPT's suggestions and the NCCN guidelines. Furthermore, ChatGPT always mentioned at least one systemic therapy aligned with NCCN's suggestions. However, there was a minimal correlation between the frequency of each cancer type and the VTQ. In summary, while ChatGPT displays promise in aligning with established oncological guidelines, its current role in assisting medical professionals and patients in making treatment decisions still needs to be defined. As the model evolves, we are hopeful that its accuracy in this area will improve, but continued research is essential to fully understand and harness its potential.

In a study by Haemmerli *et al.*,[13] the capability of ChatGPT was explored in the context of central nervous system tumor decision-making, specifically for glioma management. Using clinical, surgical, imaging, and immunopathological data from ten randomly chosen glioma patients discussed in a tumor board, ChatGPT's recommendations were compared with those of seven central nervous system tumor experts. While most patients had glioblastomas, findings revealed that ChatGPT's diagnostic accuracy was limited, with a notable discrepancy in glioma classifications. However, it demonstrated competence in recommending adjuvant treatments, aligning closely with expert opinions. Despite its limitations, ChatGPT shows potential as a supplementary tool in oncological decision-making, particularly in settings with constrained expert resources.

In a study on the effectiveness of ChatGPT in offering cancer treatment advice, Chen *et al.*[14] scrutinized the model's alignment with the NCCN guidelines for breast, prostate, and lung cancer treatments. Through four diverse prompt templates, the study assessed if the mode of questioning influenced the model's responses. While ChatGPT's recommendations aligned with NCCN's guidelines in 98% of the prompts, 34.3% of these recommendations also presented information that needed to be more in sync with the NCCN guidelines. The study concluded that, despite its potential, ChatGPT's performance in consistently delivering reliable cancer treatment advice was unsatisfactory. Consequently, patients and medical professionals must exercise caution when relying on ChatGPT and similar tools for educational purposes.

2.1. Challenges associated with LLMs as a decision-support tool in cancer care

While integrating LLMs like ChatGPT into oncology shows promise, particularly in decision support for cancer treatment, it also presents several critical challenges, as discussed in the previous section. These challenges must be addressed to ensure LLMs' safe and effective use in high-stakes medical environments. First, the issue of accuracy and precision in LLMs is a significant concern. For instance, in a study by Haemmerli *et al.*[13] on glioma therapy, ChatGPT demonstrated limitations in accurately classifying glioma types. Similarly, the study by Lukac *et al.*[7] revealed errors in patient-specific therapy suggestions, such as misidentifying patients for trastuzumab therapy. These inaccuracies highlight the risk of potential misdiagnoses or inappropriate treatment recommendations, which could have profound implications for patient care.

Another challenge is the capacity of LLMs to consider the comprehensive clinical picture, including patient functional status, which is often a nuanced judgment call made by experienced physicians. ChatGPT's moderate performance in this area, as seen in Haemmerli *et al.*,[13] indicates a gap between current LLM capabilities and the complex decision-making processes in medical practice. Furthermore, the integration of LLMs into existing medical workflows raises concerns. For example, Gebrael *el al.*[8] study on triage in metastatic prostate cancer showed that while ChatGPT had high sensitivity, its low specificity for discharges could lead to operational inefficiencies. Integrating LLMs within health-care systems also poses challenges in data privacy, interoperability, and the need for robust IT infrastructure.

Finally, the role of LLMs in patient education and communication is not without limitations. Inconsistencies in ChatGPT's responses to breast cancer prevention and screening demonstrated by Haver *et al.*[10] This inconsistency highlights the importance of human oversight in verifying the information provided by LLMs, to ensure it aligns with established medical guidelines and practices. In summary, while LLMs present exciting opportunities for enhancing cancer care, their current limitations in accuracy, comprehensive clinical assessment, integration into existing systems, and patient education necessitate a cautious and critical approach. These models should be viewed as supplementary tools that augment, rather than replace, the expertise of medical professionals. Continuous evaluation, refinement, and ethical consideration are essential to harness the full potential of LLMs in oncology.

3. Skin care (dermatology)

Our skin is a barrier against external threats such as viruses, bacteria, and other harmful organisms. Dermatology is the branch of medicine dealing with skin diseases. There has been a surge in cases related to skin diseases in the past years, affecting people of all ages.[15] Common skin-related diseases include acne, alopecia, bacterial skin infections, decubitus ulcers, fungal skin diseases, pruritus, and psoriasis.[16] Traditional dermatology diagnosis is based on a visual inspection of skin features and subjective evaluation by a dermatologist.[17] The realm of dermatology diagnosis faces several significant challenges. First, accurately interpreting skin disease imagery is complex due to the wide variety of skin conditions and their subtle visual differences. This task requires a high level of expertise, by dermatologists obviously in shortage, especially in remote or underserved areas. Finally, creating patient-friendly diagnostic reports is another hurdle because preparing reports that are detailed yet understandable to non-specialists is a time-consuming and labor-intensive endeavor for dermatologists.

In addressing the above challenges in dermatological diagnostics, Zhou *et al.*[18] introduced SkinGPT-4, an innovative interactive dermatology diagnostic system underpinned by an advanced visual LLM. This study was mainly focused on tackling the prevalent issues in dermatology, such as the shortage of specialized medical professionals in remote areas, the intricacies involved in interpreting skin disease images accurately, and the demanding nature of creating patient-friendly diagnostic reports. SkinGPT-4, utilizing a refined version of MiniGPT-4, trained on an extensive dataset that included 52,929 images of skin diseases, both from public domains and proprietary sources, along with detailed clinical concepts and doctors' notes. This comprehensive training on skin-related disease images endowed SkinGPT-4 to articulate medical features in skin disease images using natural language and make precise diagnoses. The functionality of SkinGPT-4 allows users to upload images of their skin conditions, after which the system autonomously analyzes these images. It identifies the characteristics and categorizes the skin conditions, performs an in-depth analysis, and provides interactive treatment recommendations. A notable aspect of SkinGPT-4 is its local deployment feature, combined with a solid commitment to maintaining user privacy, making it a viable option for patients seeking accurate dermatological assessments. To ascertain the efficacy of SkinGPT-4, the study conducted a series of quantitative evaluations on 150 real-life dermatological cases. Certified dermatologists independently reviewed these cases to validate the diagnoses provided by SkinGPT-4. Among the 150 cases, a commendable 78.76% of the diagnoses rendered by SkinGPT-4 were validated as either accurate or relevant by the dermatologists, breaking down into 73.13% that firmly aligned and another 5.63% that agreed. The outcomes of this evaluation underscored the accuracy of SkinGPT-4 in diagnosing skin diseases. While SkinGPT-4 is not positioned as a replacement for professional medical consultation, its contribution to enhancing patient comprehension of medical conditions, improving communication between patients and doctors, expediting dermatologists' diagnostic processes, and potentially fostering human-centered care and health-care equity in underdeveloped regions is significant.

3.1. Challenges associated with utilizing LLMs in dermatology

The introduction of SkinGPT-4 by Zhou *et al.*[18] marks a significant advancement in dermatological diagnostics, addressing challenges such as dermatologist shortage, and simplifying skin disease image interpretation and patient-friendly report generation. Despite its innovative approach and the training on an extensive dataset to articulate medical features in skin images, there are inherent challenges. Several challenges associated with deploying SkinGPT-4 include ensuring consistent diagnostic accuracy across various skin conditions, safeguarding patient privacy while managing sensitive health data, and integrating the technology seamlessly into existing healthcare systems. In addition, despite SkinGPT-4's high diagnostic accuracy, continuous human oversight in medical diagnosis and treatment planning remains critical to complement the AI's capabilities with professional medical judgment and ensure optimal patient care outcomes. In addition, advancements might focus on developing models that can adapt to new, emerging skin conditions and leveraging telemedicine to extend dermatological care to remote areas, thus promoting health-care equity.

4. Neurodegenerative disorders

Neurodegenerative disorders are characterized by the gradual deterioration of specific neuron groups, differing from the non-progressive neuron loss seen in metabolic or toxic conditions. These diseases are categorized by their primary symptoms (such as dementia, parkinsonism, or motor neuron disease), the location of neurodegeneration within the brain (including frontotemporal degenerations, extrapyramidal disorders, or spinocerebellar degenerations), or the underlying molecular abnormalities.[19] Dementia is a broad category of brain diseases that cause a long-term and often gradual decrease in the ability to think and remember, affecting daily functioning. Alzheimer's disease (AD) is the most common cause of dementia, characterized by memory loss, language problems, and unpredictable behavior.

LLM such as Google Bard and ChatGPT have emerged as valuable tools for predicting neurodegenerative disorders. A study by Koga *et al.*[20] evaluated these models' predictive accuracy using cases from Mayo Clinic conferences. The researchers extracted 25 cases of neurodegenerative disorders, from among the cases in the Mayo Clinic brain clinicopathological conferences, as their sample pool. These clinical summaries were then utilized for training and testing the models. The diagnoses offered by each model were compared against the official diagnosis provided by medical professionals. Findings from the study highlighted that ChatGPT-3.5 aligned with 32% of all the physician-made diagnoses, Google Bard with 40%, and ChatGPT-4 with 52%. When assessing the accuracy of these diagnostic predictions, ChatGPT-3.5 and Google Bard both achieved a commendable score of 76%, while ChatGPT-4 led the pack with an impressive accuracy rate of 84%. The evident proficiency exhibited by LLMs, specifically ChatGPT and Google Bard, highlights their considerable potential in revolutionizing diagnostic processes in neurodegenerative disorders.

A study conducted by Agbavor and Liang[21] explored the use of GPT-3-generated text embeddings to predict dementia, utilizing data from the ADReSSo Challenge (Alzheimer's Dementia Recognition through Spontaneous Speech *only* challenge),[22] which focuses on identifying cognitive impairment through spontaneous speech. The author proposed using the model to identify individuals with dementia against healthy individuals as controls. Using the 237 speech recordings derived from the ADReSSO Challenge, the authors used a 70/30 split and obtained 71 data samples as the testing set and 166 as the training set. In the training set, 87 individuals had AD, and 79 were healthy controls. GPT-3 was innovatively used for embedding the transcribed speech texts. Then, the model extracts the acoustic features such as temporal analysis (periodicity of speech, pause rate, phonation rate, etc.) and speech production (vocal quality, articulation, prosody, etc.). These features serve as the input for the classification model used in AD prediction. GPT-3 embeddings are then compared with BERT and traditional acoustic features. The findings reveal that text embeddings outperform traditional acoustic methods and compare well with fine-tuned models such as BERT. This suggests that GPT-3's text embeddings offer a promising approach for early dementia diagnosis.

Another study conducted by Mao *et al.*[23] outlines developing and applying a deep learning framework utilizing the BERT model for predicting the progression of an array of diseases ranging from mild cognitive impairment (MCI) to AD using unstructured electronic health records (EHR). The study cataloged 3657 MCI-diagnosed patients and their clinical notes from Northwestern Medicine Enterprise Data Warehouse (NMEDW) between 2000 and 2020, using only their initial MCI diagnosis notes for analysis. These notes underwent de-identification, cleaning, and segmentation before training an AD-specific BERT model (AD-BERT). AD-BERT transformed patient note sections into vector forms, which were analyzed by a fully connected network to predict MCI-to-AD progression. For validation, a similar methodology was applied to 2,563 MCI patients from Weill Cornell Medicine (WCM). AD-BERT outperformed seven baseline models, showing superior accuracy in both patient groups, evidenced by its area under the curve (AUC) and F1 scores.

In the diagnosis of complex conditions like AD, medical professionals use a variety of data such as images, patient demographics, genetic profiles, medication history, cognitive assessments, and speech data. Some of the recent studies have proposed multi-modal AD diagnosis or prediction methods leveraging the popular pre-trained LLM to add text data sources, in addition to images and other data types.[24-26]

4.1. Challenges associated with LLMs in neurodegenerative disorders

Utilizing LLMs in diagnosing and managing neurodegenerative disorders such as dementia and AD presents several challenges. First, the complexity and variability of these conditions require highly accurate and deep understanding, which LLMs may not always provide due to limitations in their training data. The ethical and privacy concerns about handling sensitive patient data pose significant hurdles. Furthermore, integrating these models into clinical workflows demands substantial validation to ensure they complement, rather than complicate, healthcare professionals' decision-making processes. Finally, there is a need for continuous updates and improvements in these models to keep pace with the latest medical research and clinical practices.

5. Dentistry

The World Health Organization reports that oral diseases impact approximately 3.5 billion individuals globally, with dental caries, periodontal diseases, and tooth loss being the most prevalent. These conditions, largely preventable and manageable with early diagnosis, have seen the application of AI methodologies in recent years, including the diagnosis of dental caries[27,28] and periodontitis.[29] Despite this, exploring LLMs in dentistry remains notably scarce, with limited studies demonstrating their practical application.

LLM-based deployment strategies within dentistry proposed by Huang *et al.*,[29] mark an emerging area of research with significant potential for advancement. To showcase the effectiveness and potential of applying LLMs in dentistry, this work introduced a framework for an automated diagnostic system utilizing multi-modal LLMs. This innovative system incorporated three distinct input modules, namely, visual, auditory, and textual data, enabling comprehensive analysis. Visual inputs, such as dental X-rays and computed tomography (CT) scans, are evaluated for anomalies using vision-language models to facilitate precise diagnostics. Audio inputs serve dual purposes: detecting voice anomalies and understanding patient narratives, which are converted to text for further analysis by LLM. To illustrate the capabilities of the multi-modal LLM AI system in dental practice, Huang *et al.*[29] proposed its application in diagnosing and planning treatment for dental caries. The process begins with inputting a tooth's X-ray into the system, where vision-language modeling is employed to detect any decay on the tooth. Once identified, the system utilizes LLM to propose a comprehensive treatment plan, articulated through seven detailed steps. These steps range from initial patient communication to scheduling follow-up appointments, highlighting a thorough approach to patient care. Despite its advanced diagnostics, the current system presents several limitations, such as failing to detect potential bone loss, which represent further research and development to enhance its effectiveness in dental diagnostics.

5.1. Challenges associated with dental care

The accuracy of LLMs like ChatGPT depends on the availability of high-quality, relevant dental data. A significant hurdle in designing and training LLMs for dental care is limited access to the dental records owned by private dental clinics and concerns over patient privacy, which hamper the access to comprehensive and most updated datasets. LLMs' development and effectiveness in dentistry must navigate these challenges, ensuring access to extensive, up-to-date information while addressing privacy and ownership issues to avoid biases and maintain data integrity.

The potential of LLMs in dental healthcare seems promising and can revolutionize how dental professionals diagnose, treat, and manage patient care today. LLMs could significantly improve diagnostic precision by leveraging the vast amounts of data available in patient records and imaging, allowing for early detection and intervention in dental conditions. Furthermore, the ability of LLMs to generate personalized treatment plans and educational materials tailored to individual patient needs could enhance the effectiveness of patient care. This personalization and the model's ability to process and analyze data swiftly could lead to more efficient and patient-centered dental health-care practices. As LLMs continue to evolve, their integration into dental healthcare is expected to deepen, offering innovative solutions to longstanding challenges and improving patient outcomes worldwide.

6. Mental health (psychiatry and psychology)

Mental health disorders, which affect millions globally, significantly reduce the life quality of individuals and their families. In the realm of psychiatry, LLMs have the potential to refine diagnostic precision, optimize treatment outcomes, and enable more tailored patient care, moving beyond traditional, subjective diagnostic approaches prone to inaccuracies. By leveraging AI to analyze extensive patient data, it is possible to uncover patterns not easily detectable by humans, thereby improving diagnosis.[28,29]

Galatzer-Levy *et al.*[30] delved into exploring the potential role of LLMs in psychiatry. Their primary investigation tool was Med-PALM 2, an LLM equipped with comprehensive medical knowledge. The model was trained and tested using a blend of clinical narratives and patient interview transcripts. The dataset encompassed expert evaluations using instruments like the 8-item Patient Health Questionnaire (PHQ-8) and the post-traumatic stress disorder (PTSD) Checklist Civilian Version (PCL-C). The study intended to gauge the severity of PTSD using the PCL-C while employing the PHQ-8 to assess depression and anxiety levels. The evaluation process involved extracting from Med-PALM 2 clinical scores, the rationale for such scores, and the model's confidence in its derived results. The gold standard for this evaluation was the DSM 5 (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition). The researchers' rigorous testing process involved the analysis of 46 clinical case studies, 115 PTSD evaluations, and 145 depression instances. These were probed using prompts to identify diagnostic information and clinical scores. The rigorous assessment also saw Med-PaLM 2 fine-tuned through many natural language applications and a substantial textual database. Notably, research-quality clinical interview transcripts were employed as inputs when assessing the model's efficacy. Med-PaLM 2 demonstrated its prowess in evaluating psychiatric states across various psychiatric conditions. Remarkably, when tasked with predicting psychiatric risk from clinician and patient narratives, the model showcased an impressive accuracy rate ranging between 80% and 84%.

Another study evaluated the performance of various LLMs, including Alpaca and its variants, FLAN-T5, GPT-3.5, and GPT-4, across different mental health prediction tasks such as mental state (depressed, stressed, or risk actions like suicide) using online text.[31] Through extensive experimentation, including zero-shot, few-shot, and instruction fine-tuning methods, it was found that instruction fine-tuning notably enhances LLMs' effectiveness across all tasks. Notably, the fine-tuned models, Mental-Alpaca and Mental-FLAN-T5, demonstrated superior performance over larger models like GPT-3.5 and GPT-4 and matched the accuracy of task-specific models.

The use of conversational agents based on LLMs for mental well-being support is growing; yet, the effects of such applications still need to be fully understood. A qualitative study by Ma *et al.*[32] of 120 Reddit posts and 2917 comments from a subreddit dedicated to mental health support apps like Replika reveals mixed outcomes. While Replika offers accessible, unbiased support that can enhance confidence and self-exploration, it may potentially exacerbate social isolation due to content moderation, consistent interactions, memory retention, and increased dependence on the app.

Following the advancements with ChatGPT, research into automated therapy using AI's latest technologies is gaining momentum. This new direction aims to shift mental health assessments from traditional rating scales to a more natural, language-based communication. The emergence of LLMS, like those powering ChatGPT and BERT, marks a significant shift in AI, potentially revolutionizing standardized psychological assessments. This evidence points toward AI's capacity to transform mental health evaluations into interactions that mirror natural human communication, pending comprehensive validation in specific application scenarios.[33]

6.1. Challenges associated with applications of LLMs for mental health

In mental health applications, LLMs face challenges like ensuring content sensitivity and safety to avoid generating inappropriate and harmful advice, maintaining accuracy and reliability to prevent misdiagnoses, and offering personalized, empathetic responses for adequate support. Data privacy and security are paramount due to the personal nature of discussions. There is also a need to prevent user over-reliance on LLMs, which might lead to a delay in seeking professional help. Ethical considerations include the impact of replacing human interactions with AI and avoiding biases. In addition, navigating regulatory compliance within mental health laws and guidelines is crucial for lawful operation.

7. Challenges other medical specialties

The integration of LLMs into medical specialties such as nephrology and gastroenterology remains in the early stages, as their full potential has yet to be realized. Current applications in these areas are sparse, highlighting opportunities for future exploration and implementation. This brief overview aims to shed light on the existing implementations of LLMs within these specific fields, indicating the nascent but promising role of advanced AI technologies in enhancing diagnostic and treatment methodologies in nephrology and gastroenterology.

7.1. Nephrology

Within the domain of nephrology, LLMs are being utilized to assist in diagnosing kidney diseases, providing treatment guidance, and monitoring renal function, as noted by Wu *et al.*[34] These LLMs facilitate the evaluation of crucial data such as laboratory results, clinical data, and medical history during the diagnostic phase. Various LLMs, including Orca Mini 13B, Stable Vicuna 13B, Falcon 7B, Koala 7B, Claude 2, and GPT-4, have found applications in treating and diagnosing kidney diseases. However, due to their unique zero-shot reasoning capabilities, GPT-4 and Claude 2 are particularly suitable for this intricate medical specialty. At present, these models are employed to respond to multiple-choice questions about nephrology. Wu *et al.*[34] incorporated questions regarding clinical backgrounds linked to 858 nephSAP multiple-choice queries collated between 2016 and 2023. When evaluating the proficiency of Claude 2 and GPT-4, performance was gauged based on the proportion of correctly answered nephrology-related nephSAP multiple-choice questions. GPT-4 demonstrated superior performance, garnering a score of 73.3%, in contrast to Claude 2, which achieved a score of 54.4%. When individual nephrology topics were examined, GPT-4 consistently outperformed its counterparts, including Claude 2, Vuna, Kaola, Orca-mini, and Falcon.

7.2. Gastroenterology

Lahat *et al.*[35] explored the capabilities of LLMS, specifically OpenAI's ChatGPT, in responding to queries within the realm of gastrointestinal health. Their evaluation employed 110 real-world questions, benchmarking ChatGPT's responses against the expert consensus of seasoned gastroenterologists. These queries spanned a spectrum of topics, from diagnostic tests and prevalent symptoms to treatments for a range of gastrointestinal issues. The source of these questions was public internet platforms. The researchers evaluated the outputs of ChatGPT on metrics such as accuracy, clarity, up-to-dateness, and efficacy, rating them on a scale from 1 to 5. These outputs were then categorized into symptoms, diagnostic tests, and treatments. ChatGPT averaged scores of 3.7 for clarity, 3.4 for accuracy, and 3.2 for efficacy in the symptom category. Diagnostic test-related queries resulted in scores of 3.7 for clarity, 3.7 for accuracy, and 3.5 for efficacy. As for treatment-related questions, the model achieved 3.9 for clarity, 3.9 for accuracy, and 3.3 for efficacy. The results indicated the substantial potential of ChatGPT in providing valuable insights within the gastrointestinal specialty.

7.3. Allergy and immunology

In allergy and immunology, LLMs, akin to their applications in dermatology, have shown promising potential. According to a study by Goktas *et al.*,[36] LLMs, specifically models like GPT-4 and Google Med-PaLM2, significantly enhance the diagnostic process within allergy and immunology disciplines. These advanced models elevate the precision of diagnosis and can tailor treatment plans to suit individual patient needs. Beyond the clinical realm, they also play a pivotal role in fostering patient engagement, ensuring patients are actively involved and informed during the treatment process. As a result, the integration of LLMs in allergy and immunology represents a paradigm shift toward more accurate, personalized, and patient-centric medical care.

8. Handling different types of data in the medical industry

This section provides an overview of how different data formats and types are handled in the medical industry when used as training data or inputs for an LLM.

8.1. Clinical notes

Clinical notes, an integral component of patient health records, have increasingly been utilized as input to LLMs in the medical domain. These notes, typically generated by health-care professionals, serve as rich patient information repositories, including their medical history, present symptoms, diagnoses, treatments, and more. Clinical notes are fed into LLMs to generate meaningful patterns, predictions, and insights. Before using these notes, they are often preprocessed to ensure they are in a format that is easily digestible for the models. This preprocessing can involve converting handwritten notes into digital formats, anonymizing patient data to maintain privacy, and structuring the data in a consistent format. LLMs can directly process these notes and produce a range of tools suited for activities such as condensing medical data, assisting in clinical decisions, and creating medical reports.[37] To utilize clinical notes in LLMs, prompts containing questions, scenarios, or comments about the note are used, such as "Assume the role of a neurologist at the Mayo Clinic brain bank clinicopathological conference." In response to the prompt, the model provides an output that aids in evaluation or diagnosis across different medical fields.[37]

8.2. X-rays/Images

X-rays are medical imaging that utilizes ionizing radiation to produce images of internal body organs. This data type may include CT scans (tomography), chest X-rays, and bone X-rays. In medicine, X-ray images can be processed by a computer-aided detection (CAD) model, which is pre-trained to derive the outputs in tensor form. These tensors are then translated into natural language, where they can be used as LLM input to generate summaries or descriptions of the X-ray images. Wang *et al.*[38] illustrated how the X-rays of exam images are handled while utilizing them with LLMs. They found that the model is fed into pre-trained CAD models to derive the output. They found that the images can be fed into pre-trained CAD models to derive the output. Then, the tensor (output) is translated into natural language. Finally, the language models are used to make final conclusions and summarize the results. The authors also established that X-ray images can be used as input in the LLM, where the images are fed into the model together with prompts to generate the image summarization or descriptive caption. The LLM supports visual question answering, where the X-ray images of the patients are fed into an image encoder (BLIP-2), where the natural language presentation is generated and embedded based on the image understanding.

Bazi *et al.*[39] proposed a transformer encoder-decoder architecture to handle the visual data when using LLM. They extracted the image features using the vision transformer (ViT) model, then used the textual encoder transformer to embed the questions, which were subsequently fed as the resulting textual and visual representations into a multi-modal decoder to generate the answers. To demonstrate how LLM handles the visual data, the authors used VQA datasets for radiology images, termed PathVQA and VQA-RAD. In decoding the radiology images, the proposed model achieved 72.97% and 8.99%, respectively, for the VQA-RAD, and 62.37% or 83.86%, respectively, for PathVQA.

8.3. Radiological reports

Radiological reports are documents from radiologists that present the findings or interpretation of medical imaging studies such as magnetic resonance imaging (MRI), X-rays, and CT scans. These data are processed as texts within the report to be input for LLMs in medicine. After data augmentation, the radiological reports are used as inputs in the LLM model. Tan *et al.*[40] collected and categorized 10,602 CT scan reports of cancer patients from a single facility into four response types: no evidence of disease, partial response, stable disease, or progressive disease. To analyze these reports, they utilized various models, including transformer models, a bidirectional LSTM model, a CNN model, and traditional machine learning approaches. Techniques such as data augmentation through sentence shuffling with consistency loss and prompt-based fine-tuning were applied to enhance the performance of the most effective models.

8.4. Speech data

Speech data, encompassing medical interviews, consultations, and patient audio interactions, serve as a valuable reservoir of information. Before being applied in LLMs, this data is converted into a textual format through automatic speech recognition (ASR) systems. Notably, converting audio data into text is accomplished using pre-trained models, such as Wav2vec 2.0, which has emerged as a leading contender in speech recognition technology. In their groundbreaking work, Agbavor and Liang[21] employed the Wav2vec2-base-960 base model, an advanced tool fine-tuned on an extensive 960-h dataset of 16 kHz speech audio. Their methodology incorporated Librosa for audio file loading and Wav2Vec2Tokenizer for the crucial task

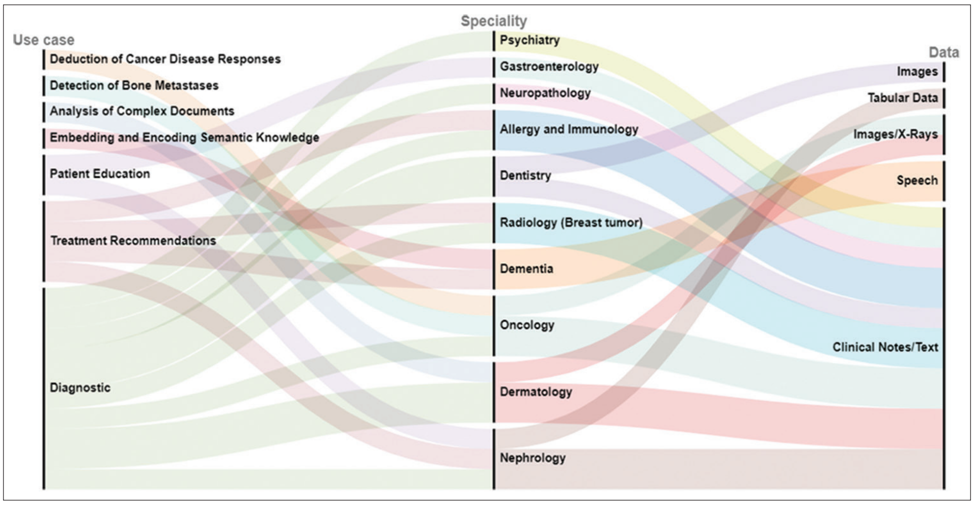


Figure 1. Visualizing large language model applications in different medical specialities with respect to input data type and medical use-case.

of waveform audio tokenization. These tokenized audio segments are inputted into the Wav2Vec2ForCTC model depending on memory capacities. This model decodes the tokens, resulting in the generation of text transcripts. Furthermore, an alternative approach to leveraging speech data in LLMs involves using open MILE, an open-source toolkit. Open MILE offers functionalities like speech classification and facilitates extracting audio features from speech or musical signals, proving its versatility in handling audio data for various applications.

8.5. Tabular data

In the medical domain, tabular data typically encompasses clinical measurements, patient records, and laboratory outcomes, arranged methodically in a matrix of rows and columns. A transformation through tabular modeling is requisite for this structured data to be effectively utilized by LLMs. The ubiquity of this tabular format in clinical and physician databases has often led to the use of tree-based models such as bagging and boosting. However, these models come with their share of limitations. Highlighting an innovative approach to this challenge, Chen *et al.*[41] presented a study employing a data set of 1479 patients undergoing immune checkpoint blockade (ICB) treatments for various cancer types. Segmenting the dataset, with 295 patients for testing and 1184 for training, they unveiled how LLMs process tabular data. Crucial to this process is serializing the feature columns into coherent sequences of natural language tokens that the LLM can interpret. This serialization can be achieved through various methods, such as prompting-based regeneration approach, using {attribute} is {value} functions, or manual serialization templates.

Furthermore, Chen *et al.*[41] introduced an advanced tabular model, ClinTaT, augmented from its original design. This refined model incorporates a continuous embedding layer harmonized with multiple distinct layers that mirror the table's continuous feature count. Continuous variables are melded with embedded categorical data for the final processing step, which is then channeled into the transformer for analysis.

9. Conclusion

LLM's applications have carved out a transformative niche in the healthcare sector. From patient engagement and education to diagnostic assistance, administrative support, and medical research, the multifaceted applications of LLMs have demonstrated their potential to optimize various facets of the medical landscape. Their expansive knowledge repositories and adeptness at understanding context and generating human-like textual responses have positioned LLMs as invaluable assets within the healthcare domain. Their integration with chatbots offers a more personalized and efficient patient experience, aiding in tasks ranging from medication clarification to mental health support. On the diagnostic front, incorporating LLMs with electronic health systems and medical imaging promises to enhance the accuracy and efficiency of diagnosis and treatment plans. LLM's capability to assist in clinical documentation, medical language translation, and medical education for patients highlights their adaptability and relevance in varied healthcare scenarios.

Despite the numerous benefits of LLMs, their practical applications in the health-care sector also underscore the importance of precision, context awareness, and ethical considerations, given the critical nature of medical decision-making. While LLMs such as ChatGPT and Med-PaLM have shown significant potential, there is an imperative for ongoing refinement, especially when handling complex or rare medical cases. As LLMs become more integrated into patient care, research addressing the ethical implications, including data privacy, the balance between automation and human intervention, and informed patient consent, will be paramount. Collaborative research exploring the fusion of LLMs with other emerging technologies, such as augmented reality or wearable health devices, can open new avenues for patient care and remote monitoring. Enhancing the LLM's contextual understanding is crucial. Future work should focus on the model's ability to consider a patient's medical history and present conditions before offering recommendations. In summary, the horizon of LLMs in healthcare is expansive and promising. As we continue to witness the convergence of technology and medicine, the collaboration of multidisciplinary teams expertise from AI, medicine, ethics, and other domains – will be integral to harnessing the full potential of LLMs in healthcare.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interest.

Author contributions

Conceptualization: All authors
Writing – original draft: All authors
Writing – review & editing: All authors
All authors contributed equally.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable

Further disclosure

The paper has been uploaded to or deposited in a preprint server (Cornell University Arxiv https://doi.org/10.48550/arXiv.2311.12882).

References

1. Min B, Ross H, Sulem E, *et al*. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput Surv.* 2023;56:1-40.

   doi: 10.1145/3605943

2. Wei J, Tay Y, Bommasani R, *et al*. *Emergent Abilities of Large Language Models*. arXiv:2206.07682 [arXiv Preprint]; 2022.

3. Brown T, Mann B, Ryder N, *et al.* Language models are few-shot learners. *Adv Neural Inform Process Syst*. 2020;33:1877-1901.

4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* 2023;29:1930-1940.

   doi: 10.1038/s41591-023-02448-8

5. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. *J Med Syst*. 2023;47:33.

   doi: 10.1007/s10916-023-01925-4

6. Sorin V, Klang E, Sklair-Levy M, *et al*. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer*. 2023;9:44.

   doi: 10.1038/s41523-023-00557-8

7. Lukac S, Dayan D, Fink V, *et al*. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. *Arch Gynecol Obstet*. 2023;308:1831-1844.

   doi: 10.1007/s00404-023-07130-5

8. Gebrael G, Sahu KK, Chigarira B, *et al*. Enhancing triage efficiency and accuracy in emergency rooms for patients with metastatic prostate cancer: A retrospective analysis of artificial intelligence-assisted triage using ChatGPT 4.0. *Cancers* (*Basel*). 2023;15:3717.

   doi: 10.3390/cancers15143717

9. Rao A, Kim J, Kamineni M, *et al*. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 Versus GPT-3.5 in a breast imaging pilot. *J Am Coll Radiol.* 2023;20:990-997.

   doi: 10.1016/j.jacr.2023.05.003

10. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology*. 2023;307:e230424.

    doi: 10.1148/radiol.230424

11. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA*. 2023;329:842-844.

    doi: 10.1001/jama.2023.1044

12. Schulte B. capacity of ChatGPT to identify guideline-based treatments for advanced solid tumors. *Cureus*.

2023;15:e37938.

doi: 10.7759/cureus.37938

13. Haemmerli J, Sveikata L, Nouri A, *et al*. ChatGPT in glioma adjuvant therapy decision making: Ready to assume the role of a doctor in the tumour board? *BMJ Health Care Inform*. 2023;30:e100775.

doi: 10.1136/bmjhci-2023-100775

14. Chen S, Kann BH, Foote MB, *et al*. Use of artificial intelligence chatbots for cancer treatment information. *JAMA Oncol*. 2023;9:1459-1462.

doi: 10.1001/jamaoncol.2023.2954

15. Yakupu A, Aimaier R, Yuan B, *et al*. The burden of skin and subcutaneous diseases: Findings from the global burden of disease study 2019. *Front Public Health*. 2023;11:1145513.

doi: 10.3389/fpubh.2023.1145513

16. Urban K, Chu S, Giesey RL, *et al*. Burden of skin disease and associated socioeconomic status in Asia: A cross-sectional analysis from the global burden of disease study 1990-2017. *JAAD Int*. 2020;2:40-50.

doi: 10.1016/j.jdin.2020.10.006

17. Burlando M, Muracchioli A, Cozzani E, Parodi A. Psoriasis, vitiligo, and biologic therapy: Case report and narrative review. *Case Rep Dermatol*. 2021;13:372-378.

doi: 10.1159/000514198

18. Zhou J, He X, Sun L, *et al*. SkinGPT-4: An interactive dermatology diagnostic system with visual large language model. 2023. medRxiv preprint.

19. Dugger BN, Dickson DW. Pathology of neurodegenerative disease. *Cold Spring Harb Perspect Biol*. 2017;9:a028035.

doi: 10.1101/cshperspect.a028035

20. Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol. 2023.

doi: 10.1111/bpa.13207

21. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. *PLOS Digit Health*. 2022;1(12):e0000168.

doi: 10.1371/journal.pdig.0000168

22. Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. *Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge*. arXiv: 210409356 [arXiv Preprint]; 2021.

23. Mao C, Xu J, Rasmussen L, *et al*. AD-BERT: Using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease. *J Biomed Inform*. 2023;14:104442.

doi: 10.1016/j.jbi.2023.104442

24. Cai H, Huang X, Liu Z, *et al*. *Exploring Multimodal Approaches for Alzheimer's Disease Detection Using Patient Speech Transcript and Audio Data*. arXiv:2307.02514 [arXiv Preprint]; 2023.

25. Feng Y, Wang J, Gu X, Xu X, Zhang M. *Large Language Models Improve Alzheimer's Disease Diagnosis Using Multi-modality Data*. arXiv:2305.19280 [arXiv Preprint]; 2023.

26. Ying Y, Yang T, Zhou H. Multimodal fusion for Alzheimer's disease recognition. *Appl Intell*. 2023;53:16029-16040.

doi: 10.1007/s10489-022-04255-z

27. Mohammad-Rahimi H, Motamedian SR, Rohban MH, *et al*. Deep learning for caries detection: A systematic review. *J Dent*. 2022;122:104115.

doi: 10.1016/j.jdent.2022.104115

28. Urban R, Haluzová, S, Strunga M, *et al*. AI-assisted CBCT data management in modern dental practice: Benefits, limitations and innovations. *Electronics*. 2023;12:1710.

doi: 10.3390/electronics12071710

29. Huang H, Zheng O, Wang D, *et al*. ChatGPT for shaping the future of dentistry: The potential of multi-modal large language model. *Int J Oral Sci*. 2023;15(1):29.

doi: 10.1038/s41368-023-00239-y

30. Galatzer-Levy IR, McDuff DN, Natarajan V, Karthikesalingam A, Malgaroli M. The capability of large language models to measure psychiatric functioning. 2023. arXiv preprint.

31. Xu X, Yao B, Dong Y, *et al*. *Leveraging Large Language Models for Mental Health Prediction via Online Text Data*. arXiv:2307.14385 [arXiv Preprint]; 2023.

32. Ma Z, Mei Y, Su Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *AMIA Annu Symp Proc*. 2024;2023:1105-1114.

33. Kjell O, Kjell K, Schwartz HA. AI-based Large Language Models are Ready to Transform Psychological Health Assessment; 2023. PsyArXiv.

34. Wu S, Koo M, Blum, L, *et al*. *A Comparative Study of Open-source Large Language Models, GPT-4 and Claude 2: Multiple-choice Test Taking in Nephrology*. arxiv: 2308.04709 [arxiv Preprint]; 2023.

35. Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: Are we there yet? *Diagnostics* (*Basel*). 2023;13:1950.

doi: 10.3390/diagnostics13111950

36. Goktas P, Karakaya G, Kalyoncu AF, Damadoglu E. Artificial intelligence Chatbots in allergy and immunology practice: Where have we been and where are we going? *J Allergy Clin*

Immunol Pract. 2023;11:2697-2700.

doi: 10.1016/j.jaip.2023.05.042

37. Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. *Nature*. 2023;620:172-180.

doi: 10.1038/s41586-023-06291-2

38. Wang S, Zhao Z., Ouyang, X., Wang Q, Shen D. ChatCAD: Interactive computer-aided diagnosis on medical image using large language models. 2023. arXiv:2302.07257.

39. Bazi Y, Al Rahhal MM, Bashmal L, Zuair M. Vision-language model for visual question answering in medical imagery. *Bioengineering*. 2023;10(3):380.

doi: 10.3390/bioengineering10030380

40. Tan RSY, Lin Q, Low GH, *et al*. Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting. *J Am Med Inform Assoc*. 2023;30:1657-1664.

doi: 10.1093/jamia/ocad133

41. Chen Z, Balan MM, Brown K. *Language Models are Few-shot Learners for Prognostic Prediction*. arXiv: 2302.12692 [arXiv Preprint]; 2023.

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

Enhancing health-care security: The role of blockchain and consensus mechanisms

Asmaul Hosna[1], Nujhat Tabassum Rahman[1], Supriya Dewanjee[1], Zulfikar Alom[1], Elmustafa Sayed Ali[2,3]*, Mohammad Abdul Azim[1], and Rashid A. Saeed[4]

[1]Department of Computer Science, Asian University for Women, Chattogram, Bangladesh

[2]Department of Electrical and Electronic Engineering, Faculty of Engineering, Red Sea University, Port Sudan, Sudan

[3]Department of Electronics Engineering, Faculty of Engineering, Sudan University of Science and Technology, Khartoum, Sudan

[4]Department of Computer Engineering, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia

***Corresponding author:**
Elmustafa Sayed Ali
(elmustafasayed@gmail.com)

Citation: Hosna A, Rahman NT, Dewanjee S, *et al*. Enhancing health-care security: The role of blockchain and consensus mechanisms. *Artif Intell Health*. 2024;1(2): 29-47. doi: 10.36922/aih.2580

Received: December 29, 2023

Accepted: February 26, 2024

Published Online: April 16, 2024

Abstract

Blockchain technology has gained prominence recently by virtue of its strong security features for clinical data. Automation of blockchain transactions enables data transactions and records, providing decentralized, secure, and dependable access. Through intelligence-sharing agreements, it can also manage member relationships without the need for a middleman or other third party. Researchers in the health-care industry using blockchain algorithms to safeguard security of data, which is properly stored, are on the rise. In addition, this technology is patient adaptive. Patients and other health-care users can now trust the technology because it prevents any third party from accessing the medical data. Many platforms intended for use in the health-care domain are emerging, including Gem Health Network and MedRec. Using blockchain in health-care protects user data and grants them full authority over their data. However, blockchain technology is also confronting challenges and limitations regarding data privacy and storage capacity. This paper explores the implementation of blockchain within health-care sector while providing an overview of this technology and the different consensus algorithms used in blockchain technology.

Keywords: Blockchain; Health-care security; Electronic health records; COVID-19 pandemic; Genetic algorithm; Consensus mechanisms

1. Introduction

Blockchain is an arrangement of a central node of control that connects distinct nodes. A peer-to-peer distribution database communication allows for safe data storage, verification, and transduction within the network. The growing utilization of blockchain technology has led to a compassionate project coordinated by the United Nations to help refugees who have lost their identity papers such as qualification documents and also to track pharmaceutical manufacturers that supply products in particular hospitals.[1] In the UK, health-care applications and validation were initialized to develop a digitized health-care system through the help of government and policy. However, the health-care applications cannot securely share data through the application with data resources.[2] Personal health-care record (PHR) technology has been introduced to safeguard and store the patients' medical records, which the patients control.

The PHR secured some confidential governance. A previous study[2] has highlighted the harmful impact of poor record sharing on patient treatment, suggesting that the patients become part of the health-care platform that requires advanced tools and capabilities. Al Mamun *et al.*[3] asserted that patients should control their medical data through electronic medical records. Since electronic medical records consist of confidential and sensitive information, the robust security system should be transparent. Blockchain ensures safe transduction and transparent medical history records – an important attribute that helps build patients' trust. For instance, based on the medical records of HIV patients, cancer patients have to endure long-duration treatment.[4] The electronic medical record services with integrated blockchain technology make the process of data storing and securing incredibly easier by storing the results of laboratory test reports, post-treatment reports, *etc.*

The Commonwealth Health Alliance took the initiative to secure the patients' electronic health records (EHRs) efficiently. Medical and health-care services are becoming increasingly important in the present environment, and they must be supplied on time, securely, and safely.[5] Disease identification has become a crucial responsibility for medical practitioners. Many viral and cardiovascular disorders, including COVID-19 and diabetes, should be diagnosed in the early phase for optimal treatment. Given its rapid dissemination, detecting coronavirus has become a critical endeavor. However, deep learning holds significant value for detecting diseases by analyzing large volumes of image data, with blockchain allowing for decentralized and secure data access.[6]

Researchers are continuously struggling to improve disease detection models. To achieve this, hospitals, testing laboratories, research centers, and other organizations can share their data and work together to improve the learning model. As the security of personal information held in hospital databases is of paramount importance concern, every party needs to commit to protecting data privacy.[7] However, accurate and efficient learning models are still required for various applications. Due to ethical and regulatory concerns about medical data privacy, data sharing among organizations are limited.

Many applications have been deployed in the battle against the COVID-19 pandemic. For example, companies such as Apple and Google have developed contact tracing applications to track COVID-19 patients.[8] However, these applications are not decentralized, so patients' data can be easily accessed, resulting in a breach of data privacy. Data in a centralized system are prone to fraud, deletion, and modifications, undermining the data integrity of these systems. Blockchain technology can assist in lowering the impacts of the COVID-19 pandemic while providing high security and ensuring that a failed attack does not occur. This is due to the decentralized feature of blockchain, and all the data and records of transactions stored in blockchain are transparent to all the network members.[1,9] Therefore, data are more reliable and trustworthy.

To make data immutable, hashing or cryptographic algorithms are applied in the blockchain, linking one block to another.[10,11] Smart contracts are also used by blockchain technology to automate business processes and resolve health-care collaborators' disputes. Blockchain technology employing intelligent contracts can be applied to the logistic supply management of COVID-19 polymerase chain reaction testing kits.[12] This can help track supplies of these kits, recognize faulty or fake kits, monitor the condition of testing kits while shipping, and allow government officials to analyze the supply and demand of testing kits in specific locations. The paper will discuss the algorithms used for blockchain security and technology in health-care sector in the aspect of protecting medical records.[13] The main contributions of this paper are illustrated as follows:

(i) *Comprehensive overview:* To provide a summary of blockchain technology and discuss its various uses and implications for the health-care industry.
(ii) *Security concerns in health care:* To outline how blockchain technology resolves important security issues in the medical field and promote it as a dependable means of protecting clinical data.
(iii) *Blockchain security algorithms:* To provide a useful insight into the technical aspects of using blockchain technology to secure medical records by discussing the algorithms used for blockchain security.

The rest of this paper is organized as follows: Section 2 summarizes related works and motivations, and section 3 introduces the background of blockchain technology. The blockchain based on consensus algorithms in health-care sector is reviewed in section 4. Section 5 encapsulates a blockchain-based EHR system for Healthcare 4.0 Applications. The taxonomy of blockchain technology in health care is discussed in section 6. In section 7, the strength of blockchain technology in the health-care sector is reviewed. Section 8 discusses blockchain technology and applications in health care. The research gap and technical limitations of blockchain in health care, and the relevant future directions, are given in sections 9 and 10, respectively. The paper is concluded in section 11.

2. Related works and motivations

Multiple studies on blockchain utilization in health-care sector have been conducted. One of the papers[14] narrates the history of blockchain development, with a focus on the technology of intelligent health-care management for assisting patients. Krishnamurthi and Shree[15] discussed several blockchain census algorithms and comparatively analyzed the algorithms. Another published study[16] presented a model to solve a confidentiality issue inherent in wearable medical devices used to monitor and care for patients, circumventing privacy intrusion and security concerns stemming from the transfer and recording of medical data. A new framework has been proposed for modified blockchain models for internet of things devices and other privacy and security features.

Yazdinejad *et al.*[17] proposed a new decentralized authentication of patients in a distributed hospital network, by leveraging the blockchain. This proposed model protects health-care networks for patients and allied health professionals. After the analysis, the results of the simulations showed that they demonstrated a high performance in ensuring confidentiality of the proposed structure among a distributed affiliated hospital network. Another study[18] expounded the different types of blockchain, such as public, private, and consortium blockchain, with elaborations on the uses of different algorithms in the health-care sector and the security purposes. Sharma *et al.*[21] proposed a framework for community interaction and developed a smartphone application to encrypt messages between researchers and research groups.

A survey conducted by Nguyen *et al.*[20] illustrated the contribution of blockchain and artificial intelligence (AI) in the health-care sector to combating COVID-19. Table 1 summarizes the most important features and contributions of the previous studies. These studies highlight the huge dependence of the next-generation health-care networks and applications on the use of the blockchain for security and user privacy.[22] Accordingly, this paper aims to comprehensively discuss the theoretical concept about the most critical blockchain issues related to the health-care sector, in addition to the impact of blockchain approaches and consensus algorithms in health-care application.

3. Blockchain technology

Blockchain is one of the most hyped disruptive innovations in recent years. It has garnered growing attention as a horizontal technology used in various sectors.[23] It is a distributed, immutable, open-source, public digital ledger distributed among network peers. It is a ledger made up of a chain of blocks. This ledger keeps a permanent record of all transactions and interactions among participants on the distributed and decentralized blockchain network. In addition, blockchain can be highly cost-effective in removing the requirement for a centralized authority to control and verify interactions and transactions between multiple users.[24] Every transaction in the blockchain is cryptographically signed and validated by all mining nodes, which keep a copy of the whole ledger made up of chained blocks of all transactions.[25] This provides unchangeable, secure, synchronized, and shareable time-stamped documents.

3.1. Types of blockchain

The three basic blockchain types are public permissionless, consortium public permission, and private blockchains.[26] They differ in terms of who has access to, writes to, and reads the data on the blockchain. Anyone can see the data in a public chain, and anyone can join and contribute to both consensus and make changes to the core software in principle. The public blockchain is commonly utilized in cryptocurrencies, and the two most popular cryptocurrencies, Bitcoin and Ethereum as the main chain, are public permissionless blockchains. Only a few specified groups of companies can monitor and participate in the consensus procedure on a consortium blockchain, which can be considered semi-centralized.[27] The private blockchain network is distributed yet often centralized. Only specific nodes can join the network, and a central authority frequently manages them.

3.2. How blockchain empowers secure data sharing in health-care system

The technologies blockchain with deep learning can improve health-care systems.[28] Utilization of blockchain technology in health-care domains helps secure data sharing and train deep learning models for diagnosing and predicting diseases. Other problems include data privacy concerns and compromised security in data flow between businesses. Therefore, the information was shared across the organization based on external and internal policies.[29] In addition, some fascinating research focuses on safe health-care data brain stimulation and biomedical and e-health data exchange for the central database built on the private blockchain by authorized users. In addition, to minimize risk, the remote patient monitoring system uses the Ethereum protocol.[30] Likewise, other authors recommended using encryption to store data from publicly accessible organizations. Several writers created a blockchain-based framework for sharing data on cloud

Table 1. Summary of related works on blockchain for securing medical data

Study	Blockchain in COVID-19 pandemic	Blockchain strength	Algorithms in health care	Taxonomy	Remarks
Mettler (2016)[14]	No	No	No	No	Describing the history of blockchain development from bitcoin and intelligent health-care management for patient guidance.
Krishnamurthi and Shree (2019)[15]	No	No	Yes	No	Presenting a comparative analysis of the algorithm, a brief overview of blockchain and challenges using algorithms.
Dwivedi *et al*. (2019)[16]	No	Yes	Yes	No	Demonstrating a blockchain-based IoT model for the security and privacy of any IoT-based remote monitoring system to protect business security.
Yazdinejad *et al*. (2020)[17]	No	No	Yes	No	Presenting a designed model for the safe data recording in a geographically diverse hospital network based on a blockchain-based approach.
Sharma *et al*. (2021)[18]	No	No	Yes	No	Proposing a cryptographic framework to create a blockchain-based secure community.
Saranya and Murugan *et al*. (2021)[19]	No	No	Yes	No	Explaining the blockchain types and uses of different algorithms in the health-care sector.
Nguyen *et al*. (2021)[20]	Yes	No	Yes	No	Combining blockchain and artificial intelligence for emergency health-care services used in the COVID-19 pandemic.

Abbreviation: IOT: Internet of things.

storage without the need for a third party.[31] Recent research has focused on real-time health-care systems' diagnosis and treatment of patient conditions.

4. Blockchain based on consensus algorithms in health-care sector

Several lists of algorithms are used in blockchain networks such as proof of work (POW), proof of stake (POS), practical byzantine fault tolerance (PBFT), recovery algorithm for fast tracking (RAFT), and delegated POS (DPOS). These algorithms are discussed in the following.

4.1. POW

POW technique required mining nodes for solving complex mathematical puzzles. After solving puzzles and node validation, the block is added to the blockchain network. The rest of the mining nodes approve the authenticity of the blocks.[32] When the miners confirm that the block is authorized, the block is attached to the blockchain by recompensing submitter mining nodes. There is lesser chance to get a false reward unless the attackers accommodate more than 50% of the mining nodes. The consensus processes-based POW provides data integrity, immutability, and reliability on the blockchain, improving the security of health-care applications.[33] Consensus techniques ensure that all participants have an accurate representation of the data by assisting in reaching an agreement on the current state of distributed database.

The POW algorithm is used in the health-care transaction. The work has traversed different consensus approaches in blockchain technology and is principally recommended for health care.[34] The sensors connect with intelligent devices and distribute the data for all possible events. Since automatic intelligent contracts are executed, the data are reliable. For instance, a sensor is connected to the human body so that the master device gathers data from the sensor to telecast it to the blockchain. Once medical data are stored on the blockchain, the POW technology guarantees that it is safe and unalterable. Since the POW is known to be decentralized. The network is more resistant to attacks because of its decentralization which offers fail-safe mechanism to protect itself against a single point of failure.[29,35] This can improve the system's overall security and dependability in the health-care industry by guarding against unauthorized access and guaranteeing the ongoing availability of vital patient data.

4.2. POS

With POS consensus mechanisms, miners are selected based on the quantity of cryptocurrency they own and are prepared to stake as collateral, thereby replacing the conventional POW mining method's intricate computational puzzle approach. In health-care applications, the medical chain is a systematic scheme of data sharing that can be executed for health-care systems using blockchain technology.[34] The incidence of attack against POS-based blockchain is lesser than that against POW-based blockchain. In POS, it is very difficult for an attacker to obtain the majority of the blockchain supply. This protects patient data integrity and increases the resilience of PoS-based health-care blockchains against threats. For health-care applications based on blockchain, switching from complicated computational problems to POS results in lower energy consumption and enables randomized validator selection, node participation incentives, performance-based rewards, and continuous work to resolve distribution issues.[36] The security, effectiveness, and dependability of health-care blockchain applications are all improved by these contributions taken together.

Moreover, using the POS consensus mechanism in the context of an EHR system indicates that health-care applications built on blockchain technology are more secure and efficient in terms of making smart decisions. Real-time modifications to patient records can be made easier with POS, which has advantages including faster transactions and less energy usage. In addition, the tasks of verifying patient records are carried out by trusted health care providers who use health care networks to ensure reliability and efficiency.[37] POS is a good option for applications where timely access to patient data is crucial and environmental concerns are present since it combines the benefits of decentralization with a fast and streamlined consensus process, which enhances the system's overall security.

4.3. DPOS

DPOS is a decentralized model with high efficiency but low consumption. There is an option to vote for creating a panel with restricted trusted parties known as witnesses.[33,38] Some users act in the reputation system. It can create blocks and add them to the blockchain. The DPOS census is cost-efficient and time-saving. Since few nodes are eligible for DPOS to be centralized, the central node can easily monitor the election process. DPOS cannot maintain all the nodes effectively, undercutting the trustworthiness in security. Nowadays, the health-care domain is undergoing advancements through the incorporation of blockchain.[34,39] The implementation of the DPOS algorithm ensures the privacy of EHRs through secure transactions. With this technology, the patients maintain control of their EHRs. The patients may share their medical records with different institutions.

Blockchain technology can ensure the privacy and security of shared data. Once a doctor updates the EHR, it is encrypted by the SHA256 hashing algorithm, and then, it is stored in a different block.[35,40] The doctor receives a unique key from the patients through mail for accessing the medical data. The DPOS algorithm secures the patient data with a trustworthy guarantee and lowers the computational time and minimizes the entire cost of processing EHRs.

4.4. PBFT

PBFT can solve the byzantine problem, as presented in a published paper.[30] A byzantine fault is a defective algorithm. Byzantine fault tolerance can ensure the safety and efficiency of the system so that hardly [(n-1)/[3] duplicate data are defective over the system in a lifetime. In medical science, PBFT algorithms create an efficient impact because several nodes are being shared and maintained by several nodes.[36,40] The fact that they hinder medical data from being disclosed or accessed by attackers significantly enhances the trustworthiness of PBFT.

4.5. RAFT

RAFT has five server nodes with three states, namely leader, follower, and candidate. Modified RAFT nodes work in a category accepting the same transitions. For instance, if a person is selected from a category assigned as a leader, he must accept clients' requests.[37,41] The leader must replicate the log to other servers and the data flow from the leader to the server. The leader's task is divided into three subtasks: leader election, leader log replication, and safety. A new leader is elected when the assigned leader fails to monitor the works. In log replication, the leader can guide and command the followers to execute changes made by the leader.[34,41] Finally, RAFT uses different commands for the same log index when the server changes the state of machine for safety concerns. Figure 1 shows the process of cluster algorithm of RAFT.

By comparing these five consensus algorithms, as shown in Figure 2, the POW algorithm stands outs as the most efficient for the health-care sector because it has a robust security system, which is the primary goal of initiating blockchain algorithms in health-care sectors. The summary of compared algorithms is shown in Table 2.

A previous study[29] provides a framework for implementing the algorithms discussed previously, where a number of computers with the same specifications were used to act as nodes for the blockchain.[38] By considering the typical framework with a computers of core I7, with the specifications of 16 GB memory size, and Window 10 operating system, the experimented POW, POS, DPOS, and PBFT algorithms with data size of 100 M/times can deliver performance depicted in Figure 3. An extended period of time is required to implement the proposed model in the system, as per empirical experiences, to compensate for the random delay between nodes.

The analysis also showed that the POW algorithm takes longer time compared to DPOS and POS algorithms. Furthermore, the PBFT consensus algorithm requires shorter time compared to other algorithms. The performance

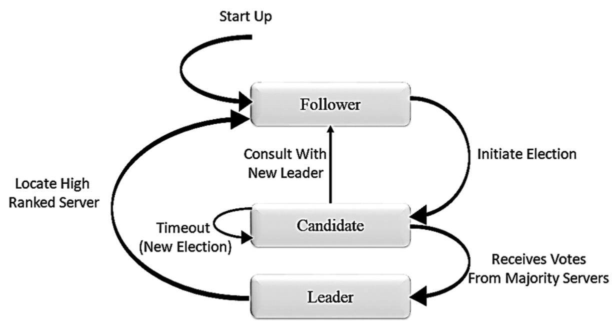


Figure 1. Recovery algorithm for fast tracking algorithm cluster diagram. Source: Schematic created by the authors.

of the PBFT algorithm is significantly different because the blockchain nodes are shared and maintained by multiple nodes, complicating the process of detecting medical data and protecting them from potential attackers.[29]

In general, the performance of these algorithms directly affects the utilization of medical data systems based on blockchain frameworks. However, depending on the kind of health-care application, the time delay during these algorithms processing will negatively impact the performance and utilization the blockchain-based health-care systems.

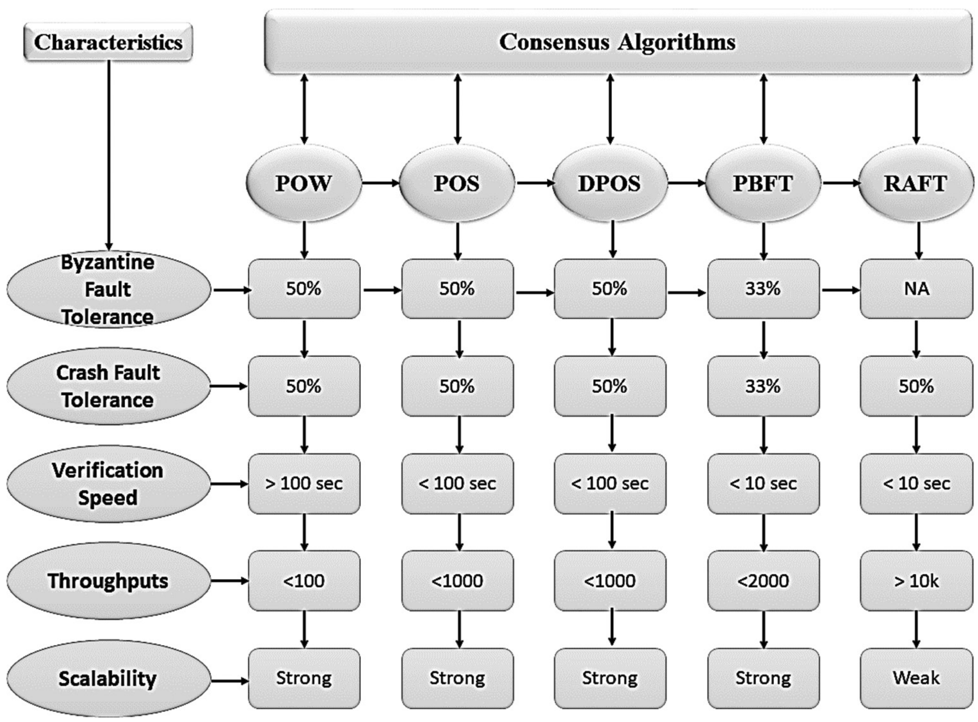


Figure 2. Performance analysis of consensus algorithms. Source: Schematic created by the authors.

Table 2. Comparative analysis of blockchain algorithm properties

Issues	Algorithms				
	POW	POS	DPOS	PBFT	RAFT
Developer	Markus Jakobsson		Developer	Markus Jakobsson	
Energy efficiency	Not enough	Limited	Limited	Yes	Efficient
Languages	C++, LLL	Michaleson	Improvised	Java	Haskell programming language
Advantages	Secure network, extensive and decentralized control over the network	Advantages	Secure network, extensive and decentralized control over the network	Advantages	Secure network, extensive and decentralized control over the network
Limitations	High consumption of electricity, not concordant with small networks			Limitations	High consumption of electricity, not concordant with small networks

Abbreviations: POW: Proof of work; POS: Proof of stake; PBFT: Practical byzantine fault tolerance; RAFT: Recovery algorithm for fast tracking; DPOS: Delegated proof of stake.

5. Blockchain-based EHR system for healthcare 4.0 applications

EHRs are medical records that can be managed and secured by a blockchain system supported by genetic algorithm and discrete wavelet transform.[32,41] The scope of a blockchain platform for industrial health care gives a new vision and future opportunities for Healthcare 4.0 Applications. The state-of-the-art focusing on the uses of blockchain with EHR in the health-care sector is summarized in Table 3.

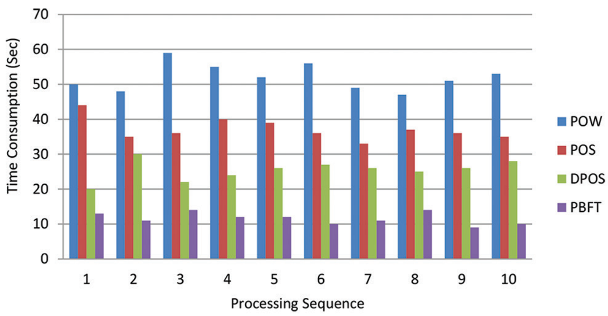


Figure 3. Consensus algorithm performance in blockchain framework. Source: Diagram created by the authors.

A study reported that EHRs of medical data consist of sensitive information and patients are permitted to share this information with health-care centers, doctors, and consultants.[15,41] EHRs are favorable for patients because they simplify the storage of laboratory reports and medicine lists and ease the appointments with attending doctors and the clinical consultations, especially for patients requiring treatment for diseases with an extended data history, such as cardiovascular disease, cancer, HIV, *etc.*[38,42] For patients who often visit different medical institutions, organizing and securing their medical history reports in the EHRs with the help of blockchain technology proves to be convenient for them. Through the data sharing features, medical research institute may collaborate with different health-care organizations under a regulated and secure data sharing environment.

Blockchain is known as a cryptographic protocol for conserving shared information records through a collection of computer networks where complete trust is not mandatory among the nodes. The implementation of blockchain in the health-care sector ensures data security for both patients and providers.[21,43] As a decentralized

Table 3. State of the art of blockchain use in health-care domain

References	Category	Components	Merits
Tanwar *et al.* (2020)[48]	Electronic health records	• Advantageous blockchain for health-care scenario • Securing and storing health-care clinical data • Data authentication for decentralized network	• A description of EHR work is presented • Transaction process in blockchain is explained • Blockchain in the health-care ecosystem is overviewed
Farouk *et al.* (2020)[24]	Electronic health records	• A brief description on the blockchain with EHR to share patients' information with the health-care centers and doctors	• Secure data sharing through excellent regulation
Hussein *et al.* (2018)[27]	Security and management of clinical records	• A brief description of blockchain networking system • Discrete wavelet transform for creating distinctive hash decrypted key • Genetic algorithm for enhancing data reliability	• Proposed method on managing and securing clinical data • Restriction on the access to the data using discrete wavelet transform algorithm • Enhancing data reliability using genetic algorithm
Wang *et al.* (2019) [46]	The SecNet	• Al-based algorithms to protect computing platforms • Smart contract algorithm i • Implementation of SecNet in medical data sharing	• Two aspects of SecNet are evaluated • Vulnerability of architecture and revenue for contributors is considered • An alternative storage model of the SecNet is proposed
Alqaralleh *et al.* (2021)[4]	Health-care diagnosis model	• An effective model for secure blockchain-enabled intelligent IoT • New health-care diagnosis model	• A data-gathering method is carried out to collect patient information using IoT devices • The GO-FFO (grasshopper with the fruit fly optimization) algorithm with elliptic curve cryptography is utilized for confidential image transmission for starters • NIS-BWT (neighborhood indexing sequence with burrow wheeler transform) approach is used to encrypt hash value • Deep belief network model is applied for diagnosing disease

Abbreviation: IOT: Internet of things.

system, the involvement of third parties is not allowed in blockchain. Thus, health-care service maintained by blockchain technology can only permit sharing of data contained within the blockchain architecture. The patients who use blockchain technology are facilitated with cost-efficient data distribution.[39,44] Moreover, the patients are privileged with an extensive network for secure health-care systems, medical data exchange through blockchain, health-care data protection, EHR facilities with attribute-based cryptosystem, and facilities for monitoring clinical emergencies. There are four stages of securing clinical data in the health-care industry:

(i) First step: At first, various health-care data, including patient's personal information and ID, are sent to the blockchain network through application programming interface (API). The current health IT system tracks and stores all the data.[43]
(ii) Second step: Blockchain technology has an internal transaction process through a smart contract. Entire transactions attached in the blockchain contain only patients' public ID rather than their personal information.
(iii) Third step: A permanent ledger is connected with the block. Thus, all sections become distinctly identifiable. The API processes queries from the health provider in a reverse manner. The database of blocks stores anonymous patient data, *e.g.*, gender, age, and illness.
(iv) Fourth step: The patient will have a private key. The health-care provider can only access the patient's information after the patient shares the private key. The data stand is restricted to people who do not have a private key.

Hussein *et al.*[27] proposed an extensive and prosperous system for handling the clinical record and information using blockchain technology. The method implements a different cryptographic technique for strong security management of sensitive clinical data and adaptability of the patients to simplified data access.[44] Discrete wavelet transform using hash function generation process was employed to boost the strength and restrict the access of data users. Moreover, genetic algorithms lower the time of transaction nodes to enhance data reliability and designate the data requests.

There are separate blocks in the blockchain network that is shaped by establishing chain events from the current block to the original block. After obtaining event details, the block broadcasts into a network.[45] Once the chain forms, the block is locked and cannot be reformed, updated, and deleted. Any exploitation of data handling policies by users in the group will prompt data tracking by data forensics team so as to secure and manage clinical records.

SecNet is an architecture proposed by Wang *et al.*,[46] combining actual big data with AI to enhance the robustness of cyber security. A large-scale Internet setting offers safe data storage, computation, and sharing. It primarily consists of three components. Blockchain-based data sharing with ownership guarantees allows trusted data exchange to create massive data in a large-scale context. In addition, AI-based safe computing systems come with more intelligent security rules, which aid in the creation of more trustworthy cyberspace. Moreover, they purchase security services through trust value exchange, a method for participants to receive financial rewards for sharing their data or service, promoting data sharing, and improving AI performance.[47] Furthermore, the authors describe a scenario of using conventional SecNet and its potentially alternative deployment method and evaluate its network security and economic revenue.

Alqaralleh *et al.*[4] developed a deep learning model for safe image transmission and diagnosis on the Internet of Medical Things environment. Data gathering, secure transactions, hash value encryption, and data classification are among the procedures included in the model.[49] The elliptic curve cryptography (ECC) is used primarily, and the hybridization of the grasshopper with the fruit fly optimization technique is used to generate the best ECC keys. The hash values are encrypted using the neighborhood indexing sequence (NIS) with burrow wheeler transform (BWT) (NIS-BWT). Finally, a deep belief network is used in the categorization process to diagnose the presence of disease. To identify the analysis of the optimal results of the proposed model, substantial experimental validation is performed, and the results are examined from many perspectives.

6. Taxonomy of blockchain technology in health care

Blockchain technology utilizes network technology with tamper-resistant data. In blockchain technology, current transactions cannot be changed. Instead, the transactions can be updated using hash values. The taxonomy of blockchain technologies in health care is illustrated in Figure 4. Different features make blockchain technology distinctive from others:

(i) *Distributed ledger:* In a distributed system, transactions are added to retrieve the system by removing failure points.
(ii) *Census mechanism:* If every verified user of the network grants a permission transaction, the transaction can be updated.
(iii) *Provenance:* The entire data history is obtainable on the blockchain network.

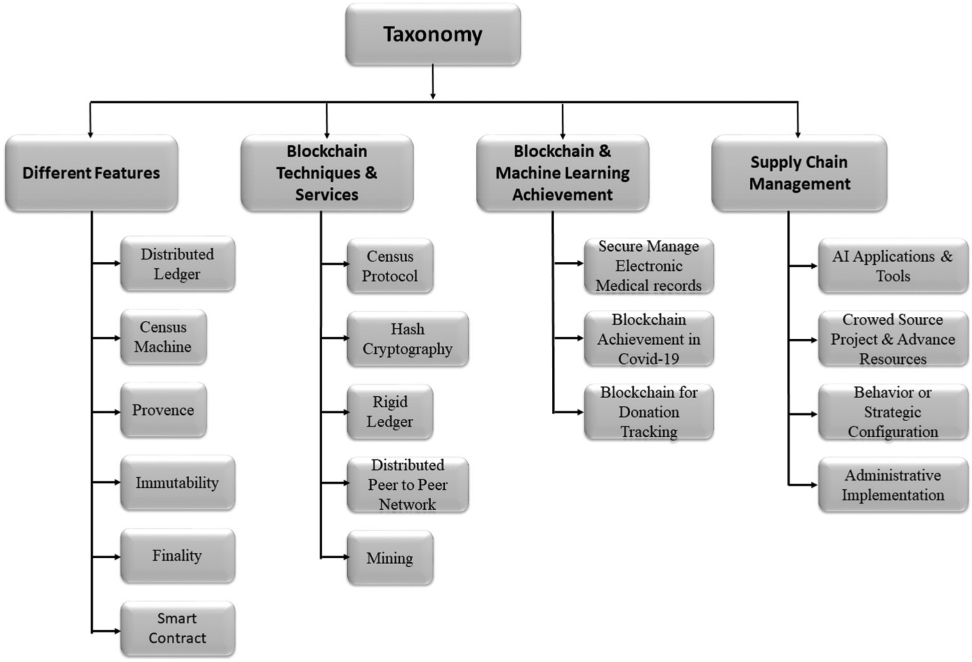


Figure 4. Taxonomy of blockchain. Source: Schematic made by the authors.

(iv) *Immutability:* Since all data are secure and trustworthy, not even a single record cannot be changed or redesigned.

(v) *Finality:* After completing the transaction, no one can change any data.

(vi) *Smart contract:* The codes are automatically executed within a specific time limit. The codes generated in the blockchain network and nodes are activated after executions. Consequently, blockchain hinders third parties from accessing the transaction, thus promising data security.[17]

6.1. Blockchain techniques and services

Blockchain incorporates various techniques and services such as Census Protocol, Hash Cryptography, Rigid Ledger, Distributed Peer-to-Peer Networking, and Mining, which are briefly explained in the following:[45,49]

(i) Census Protocol: There is a substantial restriction in allowing transactions. Through the Census Protocol, only specific users have access to the network assigned to grant permissions for transactions.

(ii) Hash Cryptography: The NSA has developed the SHA256 hash with 64 characters to add transactions used by blockchain. Hash algorithms have several uses, such as one-way cryptography, quick computation, avalanche effect, and inevitable combat impact.

(iii) Rigid Ledger: It is not possible to delete or modify because the blockchain network is stored and recorded.

(iv) Distributed peer-to-peer network: The data are updated and distributed through the network and distributed to different users.

(v) Mining: Miner helps attain the hash values in the network. The hash values can be easily computed for acquiring the award.

6.2. Blockchain and machine learning achievements

Blockchain has become a hot research topic since its inception. The concept of blockchain was first exemplified in digital currency, for example, Bitcoin in 2008. It has brought tremendous changes in health-care sector owing to its data obscurity, stability, and propagation features.

6.2.1. Safe management of electronic medical records

The accessibility and management of medical data in electronic medical records are not completely protected from any risks. The security and confidentiality of patients' confidential information, such as disease reports, medical history, and personal information, are not guaranteed. The solution to this hurdle is combining the interplanetary file system framework for electronic medical records in the health-care industry.[40,50] Inter Planetary file system (IPFS) allocates a peer-to-peer storage structure for reserving and accessing the encrypted huge volume of electronic medical records while needed. If any file needs to be deleted from version-control history, IPFS accumulates files with content address hash from a distributed hash table. IPFS uploads the hash value of the data as an alternative to keeping all medical data. IPFS makes a distinctive content address for storing and retrieving the data.[3,50]

6.2.2. Blockchain achievements during COVID-19 pandemic

Blockchain ensures that all databases are synchronized, secure, and verified. Nowadays, researchers and health-care professionals leverage blockchain technology to curb the spread of COVID-19 pandemic and create alerts about future pandemics. There are several blockchain-based practices applied in realm of health care during the COVID-19 pandemic, including tracking of infectious disease outbreaks. By virtue of its ability in safeguarding data security, blockchain can also efficiently keep track of the public health data regarding infectious disease such as COVID-19.[41] Blockchain also assists with the accurate delivery of responses and helps with treatment decision-making soon after the early detection of symptoms so as to curb the spread of pandemic. Moreover, it guides health administration to keep track of the viral activity and suspected COVID-19 cases.[51]

6.2.3. Donation tracking

Blockchain technology can be applied to track donations. It notifies the donors of any exigencies requiring an urgent inflow of funds and, most importantly, the receipt of their monetary contributions.

6.3. Management of medical supply chains

With the help of blockchain technology, medical supply chains in different industries can be properly managed, through a series of procedures involving records collection, demands tracking, and product supply during the pandemic. It also keeps track of the usage of tools and instruments by doctors and patients in a bid to prevent the inadvertent use of contaminated items.[42,51] Proper guidelines governed by several AI sectors pertaining to data security have been introduced to fight against COVID-19 and any other pandemics in future. These categories are given in Table 4**,** which describes the AI technologies used in medical supply chain management.

7. Strength of blockchain technology in health-care sector

The intrinsic properties of blockchain technology are highly compatible with applications in the health-care sector. The strength of blockchain technology contributes to various applications in the health-care sector, as shown in Figure 5. There are clear parallels between blockchain technology and the essential requirements of the current health-care infrastructure.[41,52] Table 5 encapsulates some characteristics of blockchain technologies that can facilitate resolution of certain hurdles facing the current health IT environment.

8. Blockchain technology and applications in health-care sector

In 2016, the National Coordinator for Health Information Technology Office requested proposals on blockchain applications in health-care sector, with a focus on data validation, auditing, and authorization. Such a move is driven by the potential obstacles laid ahead of the incorporation of blockchain technology in health-care domain, such as privacy concerns, compliance with regulatory requirements, and technical issues with data storage and distribution, even though this technology enables storage of complete health-care records of an individual as a blockchain use case.[42,52] Blockchain technology offers numerous opportunities in health-care sector for the secure sharing and storage of patients' data and medical records, and coupled with consensus methods, provides effective schemes in solving security issues in the health-care industry in recent real-world applications. Leveraging medical chain is one prominent example of using blockchain technology to protect EHRs and give people control over their personal health information. The emphasis on decentralization is consistent with maintaining the confidentiality and integrity of patient records, even though the precise consensus procedure is not usually mentioned.[44,52]

Several security companies like Hashed Health in the USA are committed to using blockchain technology to

Table 4. Management of medical supply chain with the aid of AI technology

Issues	Sector 1	Sector 2	Sector 3	Sector 4
Description	Application of AI with AI tools	Crowd source project	Description	Application of AI with AI tools
AI contribution	Develop systems for drugs and vaccines against COVID-19; enhance diagnosis and improve public health	Ensure data security (pandemic situations)	AI contribution	Develop systems for drugs and vaccines against COVID-19; enhance diagnosis and improve public health
Implementation examples	Quick diagnosis of COVID-19 using medical images (Mexico, Singapore)	Cognitive impact of COVID-19 (USA); COVID-19 symptom study (UK)	IEEE declaration for ethical implementation of AI system	Implementation examples

Table 5. List of the characteristics of blockchain technology and their descriptions

Characteristics	Description
Decentralization	• Blockchain technology is a serialized data structure used to establish a decentralized ledger. • Decentralization allows parties to transact data in the health-care sector without involving a third party, reducing financial bias and fraud.
Trustlessness	Payments are made only when the balance is available on the blockchain, a feature that is essential to secure financial balances.
User-centricity	• The user-centricity attribute of blockchain technology ensures patients in control of their personal financial data. • Blockchain allows the patient to become the key mediator in distributing his or her medical data. • Patient or family member must expressly grant the provider access to the patient's medical record governed by the patient's private key signature for every new medical interaction. • Every access to patient's data is recorded in the immutable transaction history of the blockchain, providing a clear record of who has accessed and edited the patient's record.
Transparency	• Every transaction data in the blockchain is publicly viewable.
Immutability	• Blockchain is impervious to data manipulation. • The immutability of the blockchain ledger means that transactions cannot be changed or removed once they have been recorded. • Blockchains serve as a data timekeeping system, allowing easy data history reporting.
Speed	• Blockchain technology helps enhance the efficiency of verification for health-care sector transactions between financial institutions.
Cost	• Blockchain technology obviates the need to pay transaction fees by removing intermediaries from the health-care transaction.

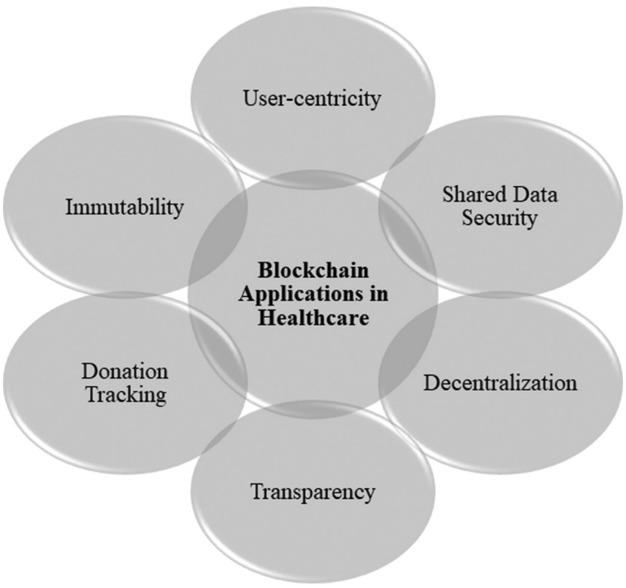


Figure 5. Summary of how blockchain technology is applied in health-care sector. Source: Schematic made by the authors.

expedite health-care transactions and minimize fraud, while enhancing operational security and efficiency. Keyless Signature Infrastructure (KSI) is utilized by other applications, including the ones developed by the Estonian E-Health Foundation, to assure the immutability of health records, prevent unauthorized adjustments, and improve overall security.[18,36] In addition, a different health-care facility uses blockchain technology to improve the security, traceability, and transparency of medical payments in an effort to lower fraud and errors in the system. According to these real-life examples, it is clear that blockchain technology can significantly add value to health-care applications.[18,44] There are other applications that have used blockchain technology in the field of health care, which are explained in the following subsections.

8.1. Blockchain for health-care management

Blockchain technology carves out a revolutionary niche in the health management sector due to the advances and benefits it brings to cloud storage of EHR data, privacy protection, *etc.*, as shown in Figure 6. With blockchain, we can improve data sharing, management, and storage. Data can then be easily shared with health-care providers. The steps of how blockchain could be used in health-care domain summarized by Khezr *et al.*[32] are given in the following:

(i) *Step 1:* While interacting with the doctors, the recent information about the patient are integrated into the medical records, serving as the primary data.
(ii) *Step 2:* EHR of the patient is shaped using the primary data collected.
(iii) *Step 3*: The control over and access to the contents embodied in EHR is granted to the EHR's owner only. Permission of the EHR's owner must be obtained for others to access EHR data.
(iv) *Steps 4, 5, and 6:* These three steps form the central part of database and cloud storage (for storing patient records) as well as data security conferred by the blockchain technology.
(v) *Step 7:* This is where health-care providers and other parties like the hospitals and care centers, collectively known as the end-users, who request access to patients' data. Records of patients' health data will be available wherever they are as they are stored and validated in the blockchain's distributed ledgers.

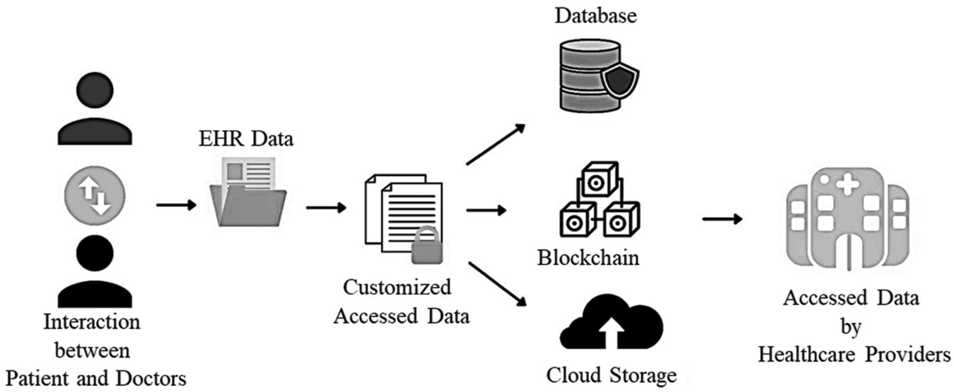


Figure 6. Health-care management using blockchain. Source: Schematic made by the authors.

8.2. Blockchain platforms used in health-care domain

Gem Health Network, Guard-time, Med-Rec, and Health-bank are some of the blockchain platforms developed for use in health-care domain. These platforms enable secure sharing of medical information with different health-care providers.

8.2.1. Gem health network

Gem Health Network is a blockchain platform developed based on the Ethereum Blockchain Framework that can allow the sharing of medical data supplied by health-care specialists. Gem Health Network merges businesses, specialists, and patients to enhance patient-centered care. This network allows medical stakeholders to have transparent access to any latest information.[44] On this platform, outdated information can be removed, thereby reducing the risk of medical negligence starting from the early stage of treatment. Medical experts can also track all the interactions between patients and their doctors.

8.2.2. Guard-time

Guard-time, a blockchain-powered data security platform based in the Netherlands, is used in Estonia to operate public health infrastructure, especially in patient identity validation. Estonian citizens are given smartcard that link their EHR data with their corresponding blockchain-based profiles. Citizens of Estonia, health-care providers, and insurance companies can acquire information about medical treatments done in Estonia through Guard-time. Updates made to the EHR are assigned with a hash and registered in the blockchain network. Thus, patients' records are immutable and are protected from malicious modifications.[45] Along with that, health-care database information, such as time and date of appointments, are also signed cryptographically in a block.

8.2.3. Med-rec

Med-Rec, built from a collaborative project between MIT Media Lab and Beth Israel Deaconess Medical Center, is a blockchain-based platform deals particularly with EHRs.[46] The non-seamless design of EHRs in managing multi-institutional and lifelong records is the prime reason for data loss as patients' data may become scattered as they move from one organization to another.[13] Med-Rec can provide its users with all of their records, which are credible, easily accessible, and most importantly, immutable. It also allows management permissions, authorization, and data sharing among health-care providers and systems using a decentralized approach.

Blockchain in the Med-Rec platform grants patients the power to authorize individuals who can access their health records. This project is tested as a proof of concept with medication data. Med-Rec was further enhanced in terms of data types as well as number of data contributors and users.[47]

8.2.4. Health-bank

Blockchain can also be applied in the area of patient-generated data. Built by a Switzerland-based digital health startup, Health-bank stands as a great example of this application. Users can store and manage their personal health-care information secured on the Health-bank platform. All the users have full control over their own data.[49] User data in Health-bank are also made available for medical research. In addition, users will be financially compensated for share their data. Blockchain implemented in health-bank allows for tracking personal patient-generated health data, which researchers can use, by means of timestamp. Users who have contributed to medical research can also be identified using blockchain.

8.3. Benefits of using blockchain technology in fighting COVID-19 pandemic

Blockchain technology offers distributed, encrypted, and secure digital transaction loggings. Blockchain technology can be employed to track the spread of coronavirus infections by tracking citizens on a global scale while maintaining patients' personal information, tracking drug trials, and tracking and maintaining records of fundraising activities and donations.[44,48] There were cases of blockchain technology used to curb the COVID-19 pandemic. Specifically, the distributed blockchain ledger technology was utilized in logging and data visualization of the coronavirus outbreaks with data derived from the Centers for Disease Control and Prevention (CDC) and the World Health Organziation.[50]

Public health blockchain consortium is another blockchain-based platform that could pinpoint communities and workplaces that have yet to be affected by coronavirus outbreaks and other pathogens before corresponding protective measures are imposed on them in a bid to prevent further spread of infectious diseases. This technology can also verify and track uninfected individuals and restrict their movements if they have visited areas affected by outbreaks.[45,51] Another example is Hyper-chain, a blockchain-based platform used in China, which facilitates donation tracking and flags the needs of COVID-19 patients to the health-care organizations and the government.

8.4. Blockchain in combating COVID-19 pandemic

Blockchain is an essential tool in the fight against COVID-19. The indispensability of the blockchain technology can be accounted for by its ability in tracking and tracing personal protective equipment (PPE).[52] During the pandemic, most countries were facing a shortage of PPEs, which are essential to prevent and control the COVID-19 infection, due to the lack of reliable and correct data about their demand and supply. A lack of transparency in the logistic supply chain management was also contributing to the prevalence of low-quality PPEs on the market. The utilization of blockchain technology can help facilitate the supply chain operations, secure PPE certificates, prevent compliance violations, and identify faulty PPEs,[18] creating an healthy atmosphere in which committers of compliance violations will be penalized, and reliable and trustworthy manufacturers will be recognized for their high-quality products.

Transactions or COVID-19 data can be recorded and made available to health-care organizations. These records of transactions will be rendered immutable, preventing alterations by any entities. Blockchain can also help with enhancing the reliability of COVID-19 analytics, reducing the incidence of fatal consequences such as COVID-19 misdiagnosis attributed to incorrect data.[43]

8.5. Blockchain for ensuring patient data privacy

Consensus procedures combined with blockchain technology offer a strong framework for protecting patient data privacy in health-care domain, resolving a number of issues commonly seen with conventional centralized methods. Ensuring the stability of patient data records by making them tamper-resistant and immutable is one of the most crucial steps in applying blockchain technology. This capability of blockchain permits the generation of an accurate and visible record of activity that documents all data transactions in the past[53] Thus, by improving data integrity and security, any unauthorized attempts to access or alter patient information can be promptly identified. Consensus techniques, which are addressed in section 4, are essential for verifying the authenticity of data supplied to the blockchain and for confirming transactions. These measures reduce the possibility of fraudulent activity and unauthorized changes to patient records by demanding network consensus. The network's trust is built through the consensus process, which improves the overall patient data security.

Blockchain networks leverage strong cryptographic algorithms to secure patient data in terms of data encryption. An extra layer of security is added by encryption, which guarantees that even in the event of illegal access, the data cannot be read without the right decryption keys. The privacy and confidentiality of patients are greatly improved by this function. By automating permissions and access controls, smart contracts self-executing algorithms with pre-established rules help protect patient privacy.[54] Patients may decide the usage and accessibility of their data through these contracts, which can be configured to manage and enforce detailed authorization procedures. Data handling in compliance with patient preferences and legal requirements is guaranteed by this automated method, which also lowers the possibility of human error.

Consensus mechanisms and blockchain technology together offer consistency, decentralization, robust encryption, access controls based on smart contracts, transparency, and improved consent management, all of which contribute to protecting patient data privacy in the health-care industry.[55] By addressing the ever-evolving issues of data security and privacy in the health-care industry, this all-encompassing strategy builds a foundation of confidence and dependability in the handling of sensitive patient data.

9. Research gap and technical limitations of blockchain in health-care sector

Blockchain technology has a positive impact on the health-care sector by facilitating the businesses of the health-care organizations. Moreover, this technology has a unique edge in securing and upgrading patient's data, in a cost-effective fashion. One of the census mechanisms, the POW, required plenty of energy to operate.[56] Due to the restriction in accessing sensitive information from the stored data, the public ledger system can be disrupted.[45] Despite its significant role in securing medical data, blockchain is fraught with limitations and challenges in technology, integration, cost, regulation, culture, energy consumption, and data privacy.

9.1. Limitations

A flood of software are currently employed in the health-care sector, but the functionalities of some of them have not fully matured and equivalent but enhanced software is constantly being created and added to this growing armada. In the aspect of integration, the blockchain technology to be applied must be compatible with the present financial technologies before their full integration.[57] Furthermore, institutions will be incurred higher initial costs due to the implementation of new technology. On a separate note, regulatory concerns surrounding blockchain technology have yet to be resolved by government agencies. One of the prominent concerns is that distributed access to the whole data set can be compromised even if the data have been encrypted and de-identified within the blockchain.[48]

The two main issues about blockchain data storage are confidentiality and scalability. Individuals who are linked on the same chain can access the data. As a result, data in the blockchain, which might contain sensitive information such as medical history and X-ray report, are vulnerable to breaches and not desirable in a decentralized platform. Storage capacity in blockchain will be highly impacted by the data breach vulnerability.[48,57] The summary of challenges facing blockchain and the guidelines to tackle each of them is shown in Table 6.

9.2. Open research issues

Several vital issues confronting the adoption of blockchain for medical applications require investigations tailored to solving security problems prevalent in the EHR systems. These open issues are iterated in four research questions:

(i) How to build servers for blockchain-based health-care systems that are amenable to blockchain protocol scalability.
(ii) How to determine the levels of authority in blockchain and safeguard the access to patient data without triggering system failure that could greatly affect access to EHR information.
(iii) How to design massive, blockchain-based globalized storage systems for large volumes of confidential health records without compromising the efficiency of the blockchain network.
(iv) When and how to integrate an approved and specialized standards formulated by global standardization institutions into blockchain-based health-care systems and into the mechanism responsible for data exchange in blockchain services.

Most recent studies present the concept of the use of blockchain in health-care domain, underscoring the important role of blockchain in transforming the health-care sector. However, one of the most important research problems surrounding the application of blockchain in health-care systems is the interoperability between different health systems following the adoption and integration of blockchain to improve security of data sharing, especially in the case of wearable devices. To investigate this aspect, Roehrs *et al.*[58] evaluated the productivity of the blockchain performance when implementing a prototype that integrates and performs medical records from different production databases.[59]

The measurement of response time, central processing unit usage, memory and disk occupation, and network usage were monitored. Figure 7 depicts the performance of blockchain in EHRs to query data and manipulate health records in a scenario containing data blocks running from 50 to 500 concurrent sessions in the network,[60] showing that there is an increase in the number of users who simultaneously access the network, measured in terms of the average load of records and the average response rate obtained.[61] These results indicate that the response time is almost equivalent despite the multiplicity and abundance of data, underlining the potential of merging open EHR standards and blockchain technologies to create an interoperable model for health data sharing with the aim of reducing the impact of various interoperability constraints.

10. Future directions

Several aspects concerning the future adoption of blockchain technology in the health-care sector should be taken into consideration:

10.1. Enhanced performance of blockchain

Platforms using blockchain technology should be technically enhanced in terms of scalability, resource consumption, network latency, throughput, *etc.* Increasing scalability and building more lightweight blockchain designs for health-care purposes are needed to make

Table 6. Challenges facing blockchain and guidelines to tackle them

Challenges	Causes	Guidelines
Security and privacy of data	• Blockchain technology is still in its early stages of development and refinement. • There is much ambiguity when it comes to designing the blockchain. • When old corporate systems and record systems are involved, integration issues arise.	• The type of data shared with and among participants must be determined since the beginning. • Prediction models that protect data privacy should be used. • It is necessary to choose a blockchain protocol – the framework that guides the structure of the blockchain and the development of applications – and use the appropriate authorization structures.
Managing storage capacity	• Storage capacity for a large amount of data is limited. • Limitations in throughput capacity and storage exist.	• A scalable and resilient blockchain solution is required. • Data storage requirements should be kept to a minimum.
Interoperability issues	• Creating blockchains from a variety of communication services is a challenge. • It is technically challenging to afford an effective interaction platform for users and for the operations of medical applications. • Gaps in communication and information sharing are obstacles.	• Evaluability should be maintained while reducing integration complicatedness. • The ease of integration should be taken into consideration with security concerns.
Decisions about blockchain governance	• Records' ownership • How is permission granted?	• New cybersecurity risks must be addressed before patients entrust a public blockchain with storing their data. • The blockchain's nodes, users, peers, and/or validators will need to be defined.
Standardization challenges	• Lack of uniformity and scalability • Lack of successful blockchain-based projects for reference	• International standardization authorities are required to formulate well-authenticated and approved standards. • The standards will be treated as guidelines for inspecting the exchanged data and as safety precautions.
Social challenges	• Concerns about blockchain adoption due to cultural and trust issues • Knowledge gap • Hesitant social adoption of technology	• Organizations are encouraged to adopt technology and join a shared network.
Inadequate universally defined standards	• No defined standards • Time- and effort-consuming implementation of standards in the health-care sector	• Universal standards will help blockchain become more adaptable. • Data format, size, and type in blockchain will be readily determined.

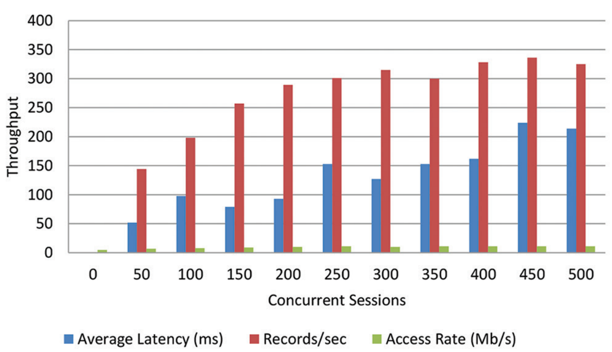


Figure 7. Interoperability performance of electronic health record systems powered by blockchain technology. Source: Graph made by the authors.

data verification and transmission of ultralow-latency information more optimized.[48,57]

10.2. Blockchain security issues

Despite the huge potential, security issues of blockchain technology that could undermine its adoption remain to be resolved. For instance, blockchain is still vulnerable to a compromise of mutual trust, by a degree of 51%, even though such a trust system is built upon with consensus mechanisms.[59] Data hackers can capitalize on this vulnerability to hijack the whole system developed with blockchain. For blockchains operating on POW mechanisms in particular, an attack with a probability of 51% may occur if one miner's hashing power is more than 50% of the hashing power in total. If a user's private key is lost, their entire blockchain will be vulnerable to tampering by other people. Since blockchain is decentralized and does not rely on third-party institutions for its operations,[54,62] it would be tough to track the whereabouts of a stolen private key and to retrieve back the stolen private key if it has been changed by the criminals. Therefore, solutions should be created to counteract the attacks and enhance blockchain security.

10.3. Reduction of resource consumption

Given the profound resource-consuming nature of POW consensus algorithm used in blockchain, a more efficient mechanism is urgently warranted. A prevailing idea of improving the existing mechanism is to develop a hybrid mechanism system of POW and POS. Further research and experiments on creating better consensus mechanisms will significantly contribute to the development of blockchain systems.[63]

10.4. Data validation and cleanup

Not all data stored in the blockchain is verified, thereby prompting smart contracts to delete some codes, although the contract address will not be removed. Furthermore, smart contracts either have the same codes or no codes at all.[59,64] In addition, most smart contracts are not published after their execution. Therefore, data cleaning and disclosure strategies must be put in place to enhance the efficiency of blockchain systems.

10.5. Future regulations

In the context of applying blockchain technology in health-care domain while safeguarding data security, more efforts should be invested in navigating and resolving the issues in the ever-changing regulatory framework. Blockchain technologies that are adherent to the current laws and regulations, including Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) requirements, should be explored.[60] To ensure compliance of these innovations with jurisdiction-specific legislations, considerations should also be given to the legal recognition and enforcement of smart contracts in health-care agreements.[65]

11. Conclusion

The integration of blockchain technology will continue to promote multifaceted advancements in the health-care industry. By comparing the algorithms utilized in blockchain technology, we found that the POW algorithm outperforms the rest. At present, several blockchain-driven platforms are already in use to store patients' medical records. These data can then be shared with medical professionals for patient-centered care and overall improvement of treatments. Patients can store their data and have full authority over who can access their data. However, there are still many flaws and challenges inherent in this technology that needs to be addressed.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no conflicts of interest.

Author contributions

Conceptualization: Asmaul Hosna, Nujhat Tabassum Rahman, Supriya Dewanjee
Writing – original draft: Asmaul Hosna, Nujhat Tabassum Rahman, Supriya Dewanjee, Zulfikar Alom
Writing – review & editing: Elmustafa Sayed Ali, Mohammad Abdul Azim, Rashid A. Saeed

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data are available from the corresponding author upon reasonable request.

References

1. Ahmad RW, Salah K, Jayaraman R, Yaqoob I, Ellahham S, Omar M. Blockchain and COVID-19 pandemic: Applications and challenges. *Cluster Comput*. 2023;26:2383-2408.

   doi: 10.1007/s10586-023-04009-7
2. Ratwani R. Electronic health records and improved patient care: Opportunities for applied psychology. *Curr Dir Psychol Sci*. 2017;26(4):359-365.

   doi: 10.1177/0963721417700691
3. Al Mamun A, Jahangir MUF, Azam S, Kaiser MS, Karim A. A Combined Framework of Interplanetary File System and Blockchain to Securely Manage Electronic Medical Records. In: *Proceedings of International Conference on Trends in Computational and Cognitive Engineering*; 2021.

   doi: 10.1007/978-981-33-4673-4_40
4. Alqaralleh BA, Vaiyapuri T, Parvathy VS, Gupta D, Khanna A, Shankar K. Blockchain-assisted secure image transmission and diagnosis model on Internet of Medical Things Environment. *Pers Ubiquitous Comput*. 2024;28:17-27.

   doi: 10.1007/s00779-021-01543-2
5. Attaran M. Blockchain technology in healthcare: Challenges and opportunities. *Int J Healthc Manage*. 2022;15(1):70-83.

   doi: 10.1080/20479700.2020.1843887
6. Abernethy A, Adams L, Barrett M, *et al*. The promise of digital health: Then, now, and the future. *NAM Perspect*. 2022;2022:1-24.

   doi: 10.31478/202206e
7. Brunese L, Mercaldo F, Reginelli A, Santone A. A blockchain based proposal for protecting healthcare systems through formal methods. *Procedia Comput Sci*. 2019;159:1787-1794.

doi: 10.1016/j.procs.2019.09.350

8. Brunese L, Mercaldo F, Reginelli A, Santone A. Lung Cancer Detection and Characterisation through Genomic and Radiomic Biomarkers. In: *2020 International Joint Conference on Neural Networks* (*IJCNN*); 2020.

   doi: 10.1109/IJCNN48605.2020.9206797

9. Ozair FF, Jamshed N, Sharma A, Aggarwal P. Ethical issues in electronic health records: A general overview. *Perspect Clin Res*. 2015;6:73-76.

   doi: 10.4103/2229-3485.153997

10. Dagher GG, Mohler J, Milojkovic M, Marella PB. Ancile: Privacy-preserving framework for access control and interoperability of electronic health records using blockchain technology. *Sustain Cities Soc*. 2018;39:283-297.

    doi: 10.1016/j.scs.2018.02.014

11. Dattani J, Sheth H. Overview of blockchain technology. *Asian J Converg Technol*. 2019;5(1):1-3.

12. Dubovitskaya A, Novotny P, Thiebes S, *et al*. Intelligent health care data management using blockchain: Current limitation and future research agenda. In: *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*. Berlin: Springer; 2019.

    doi: 10.1007/978-3-030-33752-0_20

13. Dubovitskaya A, Xu Z, Ryu S, Schumacher M, Wang F. Secure and trustable electronic medical records sharing using blockchain. In: *AMIA Annual Symposium Proceedings*. Vol. 2017. Bethesda: American Medical Informatics Association; 2017. p. 650.

14. Mettler M. Blockchain Technology in Healthcare: The Revolution Starts Here. In: *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services* (*Healthcom*); 2016. p. 1-3.

    doi: 10.1109/HealthCom.2016.7749510

15. Krishnamurthi R, Shree T. A brief analysis of blockchain algorithms and its challenges. In: Information Resources Management Association, editor. *Research Anthology on Blockchain Technology in Business, Healthcare, Education, and Government*. Hershey, PA: IGI Global; 2021. p. 23-39.

    doi: 10.4018/978-1-7998-5351-0.ch002

16. Dwivedi AD, Srivastava G, Dhar S, Singh R. A decentralized privacy-preserving healthcare blockchain for iot. *Sensors* (*Basel*). 2019;19(2):326.

    doi: 10.3390/s19020326

17. Yazdinejad A, Srivastava G, Parizi RM, Dehghantanha A, Choo KKR, Aledhari M. Decentralized authentication of distributed patients in hospital networks using blockchain. *IEEE J Biomed Health Inform*. 2020;24(8):2146-2156.

    doi: 10.1109/JBHI.2020.2969648

18. Sharma R, Wazid M, Gope P. A blockchain based secure communication framework for community interaction. *J Inf Secur Appl*. 2021;58:102790.

    doi: 10.1016/j.jisa.2021.102790

19. Saranya R, Murugan A. A systematic review of enabling blockchain in healthcare system: Analysis, current status, challenges and future direction. *Mater Today Proc*. 2021;80:3010-3015.

    doi: 10.1016/j.matpr.2021.07.105

20. Nguyen DC, Ding M, Pathirana PN, Seneviratne A. Blockchain and AI-based solutions to combat coronavirus (COVID-19)-like epidemics: A survey. *IEEE Access*. 2021;9:95730-95753.

    doi: 10.1109/ACCESS.2021.3093633

21. Sharma A, Bahl S, Bagha AK, Javaid M, Shukla DK, Haleem A. Blockchain technology and its applications to combat COVID-19 pandemic. *Res Biomed Eng*. 2020;38:173-180.

    doi: 10.1007/s42600-020-00106-3

22. Ekblaw A, Azaria A, Halamka JD, Lippman A. A Case Study for Blockchain in Healthcare: "MedRec" Prototype for Electronic Health Recordss and Medical Research Data. In: *Proceedings of IEEE Open and Big Data Conference*; 2016.

23. Ezzine I, Benhlima L. Technology against COVID-19 A Blockchain-based Framework for Data Quality. In: *6th IEEE Congress on Information Science and Technology* (*CiSt*); 2020. p. 84-89.

    doi: 10.1109/CiSt49399.2021.9357200

24. Farouk A, Alahmadi A, Ghose S, Mashatan A. Blockchain platform for industrial healthcare: Vision and future opportunities. *Comput Commun*. 2020l;154:223-235.

    doi: 10.1016/j.comcom.2020.02.058

25. Gourisetti SNG, Mylrea M, Patangia H. Evaluation and demonstration of blockchain applicability framework. *IEEE Trans Eng Manag*. 2019;67(4):1142-1156.

    doi: 10.1109/TEM.2019.2928280

26. Gupta R, Kumari A, Tanwar S. Fusion of blockchain and artificial intelligence for secure drone networking underlying 5G communications. *Trans Emerg Telecomm Technol*. 2021;32(1):1-20.

    doi: 10.1002/ett.4176

27. Hussein AF, ArunKumar N, Ramirez-Gonzalez G, Abdulhay E, Tavares JMR, de Albuquerque VHC. A medical records managing and securing blockchain based system supported by a genetic algorithm and discrete wavelet transform. *Cogn Syst Res*. 2018;52:1-11.

    doi: 10.1016/j.cogsys.2018.05.004

28. Ivan D. Moving toward a blockchain-based method for the secure storage of patient records. In: *ONC/NIST Use*

of Blockchain for Healthcare and Research Workshop. Gaithersburg, Maryland, United States: ONC/NIST; 2016.

29. Jayaram R, Prabakaran S. Onboard disease prediction and rehabilitation monitoring on secure edge-cloud integrated privacy preserving healthcare system. *Egypt Inform J*. 2020;22:401-410.

    doi: 10.1016/j.eij.2020.12.003

30. Jia Q. Research on medical system based on blockchain technology. *Medicine* (*Baltimore*). 2021;100(16):e25625.

    doi: 10.1097/MD.0000000000025625

31. Khan FA, Asif M, Ahmad A, Alharbi M, Aljuaid H. Blockchain technology, improvement suggestions, security challenges on smart grid and its application in healthcare for sustainable development. *Sustain Cities Soc*. 2020;55:102018.

    doi: 10.1016/j.scs.2020.102018

32. Khezr S, Moniruzzaman M, Yassine A, Benlamri R. Blockchain technology in healthcare: A comprehensive review and directions for future research. *Appl Sci*. 2019;9(9):1736.

    doi: 10.3390/app9091736

33. Kim D, Doh I, Chae K. Improved Raft Algorithm Exploiting Federated Learning for Private Blockchain Performance Enhancement. In: *2021 International Conference on Information Networking* (*ICOIN*); 2021.

    doi: 10.1109/ICOIN50884.2021.9333932

34. Kumar A, Kumar Sharma D, Nayyar A, Singh S, Yoon B. Lightweight proof of game (LPoG): A proof of work (pow)'s extended lightweight consensus algorithm for wearable kidneys. *Sensors* (*Basel*). 2020;20(10):2868.

    doi: 10.3390/s20102868

35. Kumar R, Wang W, Kumar J, *et al*. An integration of blockchain and AI for secure data sharing and detection of CT images for the hospitals. *Comput Med Imaging Graph*. 2020;87:101812.

    doi: 10.1016/j.compmedimag.2020.101812

36. Leeming G, Cunningham J, Ainsworth J. A ledger of me: Personalizing healthcare using blockchain technology. *Front Med* (*Lausanne*). 2019;6:171.

    doi: 10.3389/fmed.2019.00171

37. Lemieux VL. Blockchain recordkeeping: A swot analysis. *Inf Manag*. 2017;51(6):20-27.

38. Li X, Jiang P, Chen T, Luo X, Wen Q. A survey on the security of blockchain systems. *Future Gener Comput Syst*. 2020;107:841-853.

    doi: 10.1016/j.future.2017.08.020

39. Liu H, Crespo RG, Martínez OS. Enhancing privacy and data security across healthcare applications using blockchain and distributed ledger concepts. *Healthcare* (*Basel*). 2020;8:243.

    doi: 10.3390/healthcare8030243

40. Liu W, Li Y, Wang X, Peng Y, She W, Tian Z. A donation tracing blockchain model using improved DPoS consensus algorithm. *Peer-to-Peer Netw Appl*. 2021;14:2789-2800.

    doi: 10.1007/s12083-021-01102-9

41. Lo SK, Xu X, Chiam YK, Lu Q. Evaluating Suitability of Applying Blockchain. In: *2017 22nd International Conference on Engineering of Complex Computer Systems* (*ICECCS*); 2017. p. 158-161.

    doi: 10.1109/ICECCS.2017.26

42. Mackey TK, Kuo TT, Gummadi B, *et al*. 'Fit-for-purpose? Challenges and opportunities for applications of blockchain technology in the future of healthcare. *BMC Med*. 2019;17(1):68.

    doi: 10.1186/s12916-019-1296-7

43. Nakagawa T, Hayashibara N. Energy efficient raft consensus algorithm. In *International Conference on Network-Based Information Systems*. Cham: Springer; 2017.

    doi: 10.1007/978-3-319-65521-5_64

44. Shen M, Zhu L, Xu K. *Blockchain: Empowering Secure Data Sharing*. Singapore: Springer; 2020.

    doi: 10.1007/978-981-15-5939-6

45. Quiané-Ruiz JA, Pinkel C, Schad J, Dittrich J. RAFTing MapReduce: Fast Recovery on the RAFT. In: *2011 IEEE 27th International Conference on Data Engineering*; 2011.

    doi: 10.1109/ICDE.2011.5767877

46. Rahmadika S, Rhee KH. Blockchain technology for providing an architecture model of decentralized personal health information. *Int J Eng Bus Manage*. 2018;10:1-12.

    doi: 10.1177/1847979018790589

47. Wang K, Dong J, Wang Y, Yin H. Securing data with blockchain and AI. *IEEE Access*. 2019;7:77981-77989.

    doi: 10.1109/ACCESS.2019.2921555

48. Rajput DS, Sharma S, Tiwari SK, Upadhyay A, Mishra A. Medical data security using blockchain and machine learning in cloud computing. In: *Mathematical Modeling and Soft Computing in Epidemiology*. Boca Raton: CRC Press; 2020.

    doi: 10.1201/9781003038399-18

49. Tanwar S, Parekh K, Evans R. Blockchain-based electronic healthcare record system for healthcare 4.0 applications. *J Inf Secur Appl*. 2020;50:102407.

    doi: 10.1016/j.jisa.2019.102407

50. Rupa C, Midhunchakkaravarthy D. Preserve Security to Medical Evidences Using Blockchain Technology. In: *2020 4th International Conference on Intelligent Computing and Control Systems* (*ICICCS*). Madurai, India: IEEE; 2020.

    doi: 10.1109/ICICCS48265.2020.9120948

51. Ahir S, Telavane D, Thomas R. The Impact of Artificial Intelligence, Blockchain, Big Data and Evolving Technologies in Coronavirus Disease-2019 (COVID-19) Curtailment. In: *2020 International Conference on Smart Electronics and Communication* (*ICOSEC*). Trichy, India: IEEE; 2020. p. 113-120.

doi: 10.1109/ICOSEC49089.2020.9215294

52. Salah K, Rehman MHU, Nizamuddin N, Al-Fuqaha A. Blockchain for AI: Review and open research challenges. *IEEE Access*. 2019;7:10127-10149.

doi: 10.1109/ACCESS.2018.2890507

53. Sethy PK, Behera SK, Ratha PK, Biswas P. Detection of Coronavirus Disease (COVID-19) Based on Deep Features and Support Vector Machine. *Preprints, electrical and electronic engineering*, v(2), 2020, pp. 1-10.

54. Reegu FA, Abas H, Gulzar Y, *et al*. Blockchain-based framework for interoperable electronic health records for an improved healthcare system. *Sustainability*. 2023;15(8):6337.

doi: 10.3390/su15086337

55. Chinnasamy P, Albakri A, Khan M, Raja AA, Kiran A, Babu JC. Smart contract-enabled secure sharing of health data for a mobile cloud-based e-health system. *Appl Sci*. 2023;13(6):3970.

doi: 10.3390/app13063970

56. Singh M, Singh A, Kim S. Blockchain: A Game Changer for Securing IoT Data. In *2018 IEEE 4th World Forum on Internet of Things* (*WF-IoT*). Singapore: IEEE; 2018.

doi: 10.1109/WF-IoT.2018.8355182

57. Singh SK, Rathore S, Park JH. Blockio intelligence: A blockchain-enabled intelligent IoT architecture with artificial intelligence. *Future Gener Comput Syst*. 2020;110:721-743.

doi: 10.1016/j.future.2019.09.002

58. Siyal AA, Junejo AZ, Zawish M, Ahmed K, Khalil A, Soursou G. Applications of blockchain technology in medicine and healthcare: Challenges and future perspectives. *Cryptography*. 2019;3(1):3.

doi: 10.3390/cryptography3010003

59. Roehrs A, da Costa CA, da Rosa Righi R, da Silva VF, Goldim JR, Schmidt DC. Analyzing the performance of a blockchain-based personal health record implementation. *J Biomed Inform*. 2019;92:103140.

doi: 10.1016/j.jbi.2019.103140

60. Zhang P, Schmidt DC, White J, Lenz G. Blockchain technology use cases in healthcare. In: *Advances in Computers*. Vol. 111. Amsterdam: Elsevier; 2018. p. 1-41.

doi: 10.1016/bs.adcom.2018.03.006

61. Zhang R, Xue R, Liu L. Security and privacy on blockchain. *ACM Comput Surv*. 2019;52(3):1-34.

doi: 10.1145/3316481

62. Chinnasamy P, Vinothini C, Arun Kumar S, Allwyn Sundarraj A, Annlin Jeba SV, Praveena V. Blockchain technology in smart-cities. In: Panda SK, Jena AK, Swain SK, Satapathy SC, editors. *Blockchain Technology: Applications and Challenges. Intelligent Systems Reference Library*. Vol. 203. Cham: Springer; 2021.

doi: 10.1007/978-3-030-69395-4_11

63. Zubaydi HD, Chong YW, Ko K, Hanshi SM, Karuppayah S. A review on the role of blockchain technology in the healthcare domain. *Electronics*. 2019;8(6):679.

doi: 10.3390/electronics8060679

64. Zamri N, Mohamad Z, Nik WN, Mohamad AHZ. Smart secure telerehabilitation apps for personalized autism home intervention using blockchain system. In: *Blockchain for 5G-Enabled IoT*. Cham: Springer; 2021. p. 377-398.

doi: 10.1007/978-3-030-67490-8_15

65. Wang F. *Building High-performance Distributed Systems with Synchronized Clocks*. Stanford, CA: Stanford University; 2019.

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

Artificial intelligence in diagnosis and monitoring of atopic dermatitis: From pixels to predictions

Pratheek Jain[1,2], Farhan Zameer[1]*, Kounaina Khan[1], Vinay Alva[1,3], Ravish Huchegowda[4], Ali Jawad Akki[5], Raghu Anjanapura Venkataramanaiah[5], Muthuchelian Krishnasamy[6], Dilip Apturkar[3], and Raghavendra Hallur Laxmanashetty[2]*

[1]PathoGutOmics Laboratory, Alva's Traditional Medicinal Archive (ATMA), Department of Ayurveda Pharmacology, Alva's Ayurveda Medical College, Moodubidire, Karnataka, India

[2]Centre for Biotechnology, Pravara Institute of Medical Sciences, Ahmednagar, Maharashtra, India

[3]Department of Surgery, Dr. Balasaheb Vikhe Patil Rural Medical College, Pravara Institute of Medical Sciences (Deemed to be University), Ahmednagar, Maharashtra, India

[4]Department of Neurochemistry, National Institute of Mental Health and Neuro Sciences (NIMHANS), Bengaluru, Karnataka, India

[5]Department of Chemistry, Faculty of Science and Technology, Bijapur Lingayat District Educational (Deemed to be University), Vijayapura, Karnataka, India

[6]Department of Biotechnology, M.G.R. College, Dr. MGR Nagar, Hosur, Tamil Nadu, India

***Corresponding authors:**
Raghavendra Hallur Laxmanashetty
(director.biotech@pmtpims.org)
Farhan Zameer
(farhanzameeruom@gmail.com)

Citation: Jain P, Zameer F, Khan K, *et al.* Artificial intelligence in diagnosis and monitoring of atopic dermatitis: From pixels to predictions. *Artif Intell Health.* 2024;1(2): 48-65. doi: 10.36922/aih.2775

Received: January 18, 2024

Accepted: March 1, 2024

Published Online: April 18, 2024

Abstract

In any ailment, the identification of the symptoms, detection, and diagnosis plays a pivotal role in treatment and therapy. However, certain diseases share similar symptoms, lacking signature key indicators, which can lead to fallacious or incorrect inferences. Skin disorders, such as pruritus, dermatitis, eczema, psoriasis, and ichthyosis, all present similar symptoms, which confound clinicians. One such commonly misunderstood condition is atopic dermatitis (AD), a chronic inflammatory skin condition characterized by its relapsing nature, which heightens the importance of diagnosis and disease monitoring for effective management. Recent strides in artificial intelligence (AI) have opened avenues for precise diagnosis and continuous monitoring of AD. This review explores and evaluates current applications of AI in the diagnosis and monitoring of individuals with AD emphasizing the need to address challenges and collaborate across intra-, inter-, trans-, and multi-disciplinary domains to maximize the benefits of AI in enhancing the precision of AD diagnosis, ultimately leading to improved patient care and satisfaction through technologically-driven biomedical tools in customized healthcare.

Keywords: Deep learning; Machine learning; Convolutional neural networks; Artificial neural network; Pruritus

1. Introduction

Artificial intelligence (AI) has greatly improved our quality of life and is now widely used in many areas, including healthcare. Examples of its applications include autonomous driving, industrial automation, and the widespread use of cell phones. Recent developments have notably bolstered the efficiency, accuracy, and productivity of AI-optimized workflows in the health-care industry. AI involves the application of sophisticated computational algorithms to simulate complex cognitive functions of humans, accomplished through learning and adaptation to gathered data. Over the past decades, there has been a marked increase in both research and application of AI in healthcare, with the potential to completely transform the sector.[1-4]

Similar to AI, the field of dermatology has experienced rapid growth owing to advancements in technology and inventions. This evolution has brought dramatic changes in the diagnosis and treatment of dermatological illnesses. Computer algorithms have proven to be invaluable tools for dermatologists, particularly in diagnosing diseases such as malignant melanoma.[5] Dermatology boasts a vast archive of clinical, dermatoscopic, and dermatopathological images, positioning it as a leader in the application of AI in medicine. Therefore, having a basic understanding of AI becomes essential for designing and evaluating medical research in this area. Consequently, investigating the potential uses of AI in dermatological practice becomes imperative.

Atopic dermatitis (AD) is a common inflammatory skin disease that affects a significant proportion of dermatology patients worldwide, estimated to affect 1 – 2% of the global population.[6] Notably, between the 1980s and the early 2000s, there was a discernible global surge in the prevalence of AD, particularly pronounced among children under the age of five, with rates ranging from 10% to 16.5%.[7,8] Individuals with AD exhibit a wide range of clinical symptoms, categorized into six distinct subtypes based on their origin. Among these, the most prevalent subtype is early-onset, early-resolving; nonetheless, recurrence is frequently observed, with less severe symptoms than the original episode.[9,10] The three main symptoms of AD – skin inflammation, compromised skin barrier function, and persistent itching-exerta negative influence on the lives of those afflicted, significantly impacting their quality of life and level of satisfaction with therapy.[11-13] Moreover, patients' adherence to treatment procedures is severely hindered by these symptoms.[14-16] *Staphylococcus aureus* colonization often arises as a result of skin barrier dysfunction, exacerbating the degradation of the barrier function.[17,18] Notably, AD is linked to systemic inflammatory conditions, such as metabolic syndrome and cardiovascular disease, despite not being as visually conspicuous as psoriasis.[19]

Given its broad spectrum of clinical symptoms, AD presents as an unexpected illness that can prove challenging to diagnose and monitor effectively. Prompt and accurate identification is indispensable for the effective management of AD. However, conventional diagnostic techniques tend to heighten variability in diagnosis and may cause delays in the initiation of therapy, relying heavily on clinical judgment. A promising avenue toward achieving objective, accurate, and rapid diagnostic procedures lies in the integration of AI-based technology, which holds the potential to revolutionize AD management. This review highlights the rationale behind the potential of AI in completely transforming AD monitoring and diagnostic procedures. While acknowledging the efficiency of AI, it also emphasizes the importance of problem-solving and fostering teamwork. These initiatives are essential in maximizing the benefit of AI in improving the precision and effectiveness of AD monitoring and diagnosis procedures.

2. Methodology

In conducting a systematic literature review on the contemporary utilization of AI in diagnosing and monitoring individuals with AD, we implemented a rigorous methodology for gathering pertinent articles from prominent databases, including PubMed, Springer, and Elsevier. The search utilized specific keywords, such as "Artificial intelligence," "Machine learning," "Deep learning," and "Atopic dermatitis." Article selection prioritized peer-reviewed studies in dermatology that specifically examined the integration of AI in the diagnosis and monitoring of AD. Articles that failed to meet these criteria or published in languages other than English were systematically excluded from the study. Following data extraction, we summarized the key findings, and the synthesized information was then carefully compiled into a comprehensive literature review, providing valuable insights into the current state of knowledge, addressing challenges, and advocating for collaboration across intra-, inter-, trans-, and multi-disciplinary domains to optimize the benefits of AI in improving the accuracy of AD diagnosis.

3. Principle of AI

AI encompasses various computational subfields, including machine learning (ML) and natural language processing (NLP), enabling computer systems to mimic human cognitive functions (Figure 1). At present, ML, where computers anticipate data without explicit programming, stands as the frontier of AI advancement. Essentially, computers "learn" from data, offering analyses without explicit guidance on trait prioritization. Dermatologists offer compelling examples, such as identifying melanomas from clinical images,[5] predicting the effectiveness of

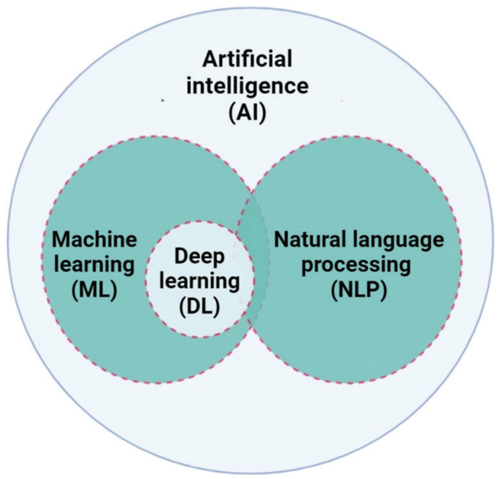


Figure 1. An overview of the principles of artificial intelligence. Artificial intelligence (AI) is a broad category of algorithms that includes subcategories such as machine learning (ML), natural language processing (NLP), and deep learning (DL).

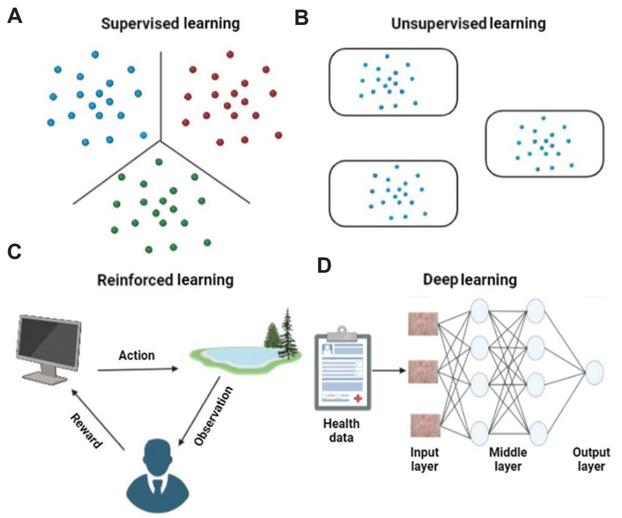


Figure 2. Several types of machine learning techniques, including supervised learning, unsupervised learning, and reinforcement learning. (A) Supervised learning involves using labeled datasets to categorize data, while (B) unsupervised learning does not use labeled datasets and instead finds patterns and relationships in the data to create categories. (C) Reinforcement learning uses iterative feedback loops to teach the algorithm. (D) Deep learning, a subset of machine learning, uses representation layers in a neural network to increase the abstraction of the data and employs techniques from all three types of machine learning.

biologic therapies for psoriasis,[20] and analyzing doctor notes in electronic health records to discern clinic visit purposes for AD.[21]

Deep learning (DL), a subset of ML, uses algorithms modeled after human neurons to discern complex patterns and relationships in data. DL permits the direct entry of raw data, unlike older ML methods that require domain expertise and human engineering to translate raw data into intelligible algorithm features.[22] For pattern recognition, the machine autonomously creates its own representations, which are arranged in a series of layers that build on one another to gradually abstract the data. Neural networks are represented by this layer architecture.[23] DL includes diverse methods, such as transformers,[24] which are adept at identifying sequential data relationships and extracting meaning, and convolutional neural networks (CNNs),[25] which are frequently used in imaging tasks. Due to its versatility and flexibility, DL is an effective tool for a wide range of applications.

Algorithms in the broad field of ML employ diverse techniques to acquire knowledge. As depicted in Figure 2, these techniques include reinforcement learning, unsupervised learning, and supervised learning. Supervised learning, the most popular ML technique, relies on labeled datasets to predict outcomes. Predictions based on unseen data are made possible by the algorithm's ability to map input data to the correct output. During the training phase, the algorithm receives both the data and the corresponding answers (ground truths) from a set of training instances, enabling it to modify its weights accordingly. Subsequently, the algorithm's performance is assessed against a different test set that it had not encountered before post-training. Logistic regression and linear regression, two of the most frequently used methods in this domain, find common applications in image-based dermatology models. On the other hand, unsupervised learning involves training a model on datasets without labels, meaning that the input lacks a known right response. The algorithm's primary goal is to discover patterns and links in the data, such as clustering related data points. In the paradigm of reinforcement learning, an algorithm referred to as the agent interacts with the environment to accomplish predetermined objectives. Based on feedback, it receives from its actions in the form of rewards or penalties, the agent modifies its behavior to optimize rewards. Reinforcement learning learns through ongoing feedback loops, distinct from the predetermined data input found in supervised and unsupervised learning methods. This three-class categorization of ML techniques offers a thorough grasp of the various techniques algorithms use to learn from data.

Drawing on concepts from linguistics, statistics, ML, and DL, a discipline of AI known as NLP aims to interpret, analyze, and generate human language.[26] This complete technique enables the processing of human language in its entirety. Within NLP, two primary subfields exist: natural language generation (NLG) and natural language understanding (NLU). NLG focuses on generating new text, while NLU is committed to understanding textual material. NLG encapsulates recent advances in large language models, exemplified by OpenAI's freely available Chat Generative Pre-trained Transformer.[27] These advancements highlight the evolving landscape of NLP and its important role in the advancement of language-related applications.

In recent years, the adoption of multimodal techniques in algorithms has surged, driven by the utilization of diverse data sources for training. Given the inherently multifaceted nature of medicine, where doctors must interpret a wide range of data, including genetic information, laboratory results, clinical notes, and radiological images, these multimodal approaches have gained prominence. The latest strides in this discipline focus on building more reliable models and algorithms by leveraging the abundance of readily available data. Noteworthy examples of these multimodal technologies include Med-Flamingo,[28] LLaVa-Med,[29] Med-PaLM Multimodal (Med-PaLM M),[30] and MiniGPT-4.[31] At the core of these technologies, they lie foundation models (FMs), which undergo training on a variety of unlabeled datasets before being adjusted for certain downstream applications.[32] One particularly intriguing aspect is the ability of these models to absorb vast amounts of information from large datasets and subsequently apply this knowledge to specific applications, including those within the medical domain. This pattern represents a dynamic movement in the direction of using multimodal techniques to improve performance in medical applications.

4. AI for the diagnosis of AD

Accurate dermatological diagnosis and treatment of AD hinge on the quantitative evaluation of the disease, which emphasizes the molecular composition of the skin using non-invasive techniques. Confocal Raman micro-spectroscopy (CRM) serves as a tool for assessing the skin's biomolecular composition. Nevertheless, deciphering complex Raman spectroscopic signals requires multivariate analysis. Dev *et al.*[33] have presented a novel approach to classifying AD from healthy individuals by combining CRM with multivariate analysis, more precisely, partial least squares discriminant analysis (PLS-DA). While the current PLS-DA classification model is designed for binary classification, there is potential to explore its applicability for multiclass categorization based on the severity of eczema illness. The ML-aided PLS-DA classification approach used in the study simplifies dimensional reduction, variable selection, and classification for Raman micro-spectroscopy data. The cross-validated PLS-DA classification model exhibits remarkable sensitivity and specificity, scoring 0.94 and 0.85, respectively. Further enhancement of categorization accuracy is feasible by concentrating on wave number bands with a variable importance in projection (VIP) score of ≥1. In addition to bolstering the model's accuracy, the VIP score facilitates the identification of important Raman spectroscopic signatures associated with proteins, lipids, and nucleic acids, which can serve as biomarkers for therapeutic and clinical evaluation of AD patients' skin health. Using CRM and multivariate analysis, this quantitative method of assessing skin inflammatory disorders such as AD offers a viable path for next-generation diagnosis, departing from the subjective scoring systems currently used in clinical practice. The presented study describes a novel diagnostic method specific for AD using CRM and multivariate analysis. This non-invasive method will provide a new approach for molecular-based evaluation of skin conditions. Nevertheless, several challenges need to be addressed, such as sample size and diversity, independent dataset validation, clinical utility assessment, CRM standardization across different laboratories, patient data privacy and informed consent ethics issues, equipment accessibility, and cost. Regulatory approval for CRM technology's widespread use is also necessary. Overcoming these issues will improve the power and generalizability of this innovative diagnostic protocol for AD. Furthermore, the application assures the availability of significant datasets and ensures the repeatability and reliability of the model.

Multiphoton tomography (MPT) has previously demonstrated its utility as a diagnostic tool in dermatology. However, MPT data analysis has remained time-consuming and operator-dependent. In a study conducted by Guimarães *et al.*,[34] the potential of using AI for diagnosing AD from MPT images was substantiated. AD system was developed to discern images containing living cells and performs subsequent diagnostics accurately and reliably, thus eliminating the need for human operators. The study has demonstrated the potential of completely harnessing MPT through a CNN-based, fully automatic method. CNNs were trained and fine-tuned using 3663 MPT images, including morphological and metabolic information from both AD patients and healthy individuals. The primary objectives were to identify live cells and diagnose AD, irrespective of the imaging layer or location. Impressively, the suggested algorithm successfully diagnosed AD in 97.0±0.2% of the images containing live cells, with a sensitivity of 0.966±0.003, specificity of 0.977±0.003, and *F*-score of 0.964±0.002. The interpretability of the algorithm was enhanced using relevance propagation through deep Taylor decomposition, generating heat maps that highlighted important details for each classification. The study exemplifies the successful integration of MPT imaging and AI for AD diagnosis, marking a substantial improvement in the field. The proposed method establishes a framework for automating the identification of skin conditions using MPT.

Activation-regulated chemokine (TARC/CCL17) and immunoglobulin E (IgE) have served as biomarkers for AD in traditional approaches over the past few decades.[35-38] Common techniques used in these investigations include regression or correlation analyses between potential biomarkers and the intensity of AD symptoms, as well as univariate research comparing AD patients to healthy controls. However, accurate diagnosing and evaluating AD solely based on a single biomarker are considered extremely challenging.[39] Recent developments have ushered in the utilization of multivariate ML techniques to uncover hidden patterns between variables and develop more reliable predictive models in a variety of studies, including those in pain research.[40,41] The combination of multiple serum biomarkers, such as TARC, IL-22, and sIL-2R, has improved the model's capability to predict eczema area and severity index (EASI) scores compared to relying on a single biomarker.[39] In addition, the correlation coefficient between the combined biomarkers and the disease severity surpasses that of the individual biomarkers. This method emphasizes the potential of combining multiple biomarkers for a thorough comprehension and prediction of AD severity.

In a recent study, Lee *et al.*[42] investigated the potential of employing a multivariate ML technique to develop a diagnostic tool and severity prediction model for patients with AD. The authors conducted phase I ML analysis, wherein they collected multivariate data, divided it into training and test sets, trained the models, estimated prediction performance, and selected and estimated features. Clinical and serological indicators[43,44] from a prior clinical study were combined. The results indicate that the classification model significantly outperformed the random permutation model, boasting an area under the curve of 0.85±0.10 and a balanced accuracy of 0.81±0.15, compared to 0.50±0.15 for the latter. Correlation analysis unveiled a significant positive association between the objective SCORing AD score (SCORAD) (r=0.53), measured and projected total SCORAD (r=0.43), and eczema area and severity index scores (r=0.58, each p<0.001). Nevertheless, no discernible relationship was observed between the measured and anticipated itch scores (r=0.21, p=0.18). The research encompassed the creation and evaluation of multivariate prediction models, as well as the identification of critical characteristics using a range of serum biomarkers. These results underscore the potential of utilizing a multivariate ML approach to reveal complex connections between clinical and serum measures in patients with mild-to-moderate AD.

Colonocytes or colonic epithelial cells have recently garnered attention for their role in host-microbial interactions. During gut dysbiosis, which is linked to a number of chronic human disorders, colonocytes influence the composition and activity of the gut microbiota.[45] The diagnosis and prognosis of AD can now be achieved through the integration and correlation analyses of gut microbiota and host gene expression.[46,47] Notable correlations have been observed, including those between IL-17 and *Streptococcus* infection in AD, and between enzyme commission genes and microbiota in inflammatory bowel illnesses.[48,49] Despite these advancements, few researchers have investigated ML prediction analysis based on the gut transcriptome and microbiota in AD. In recent work, Jiang *et al.*[50] developed an ML classifier for precise and automated AD detection by utilizing the transcriptome of gut epithelial colonocytes and gut microbiota data (Figure 3). With an average F1-score of 0.84, the classifier demonstrated accurate discrimination and successfully predicted the risk of AD. It was trained on data from 161 participants, including both AD patients and healthy controls. The research identified three genes and three bacteria that are either directly or indirectly linked to AD, as well asa combination of 35 genes and 50 microbiome traits predictive for AD. These results suggest that the discovered genes and microbiota traits may provide fresh biological perspectives and serve as useful biomarkers for early detection of AD. However, replication studies with different populations are necessary to validate these findings. The study represents a major step toward the construction of an ML classifier for accurate and automated AD diagnosis, utilizing gut microbiota and transcriptome data from gut epithelial colonocytes. The robust ML pipeline used in the study, which comprises thorough procedures such as feature selection, model selection, cross-validation, classification, and follow-up statistical assessments, enables accurate distinction based on omics data (Figure 4).

Dautović *et al.*[51] developed an artificial neural network (ANN) specifically designed for the automated diagnosis of AD, aiming to facilitate the diagnosis process. The network uses a feed-forward ANN with nine input parameters and one output parameter for classification. After evaluating various configurations, the final design of the expert system chose a neural network with 15 neurons in a hidden layer based on training results. Demonstrating impressive sensitivity at 95.62% and accuracy at 94.44%, the ANN exhibits excellent performance in distinguishing AD from other skin disorders. However, it is imperative to recognize that despite its high sensitivity, the comparatively lower

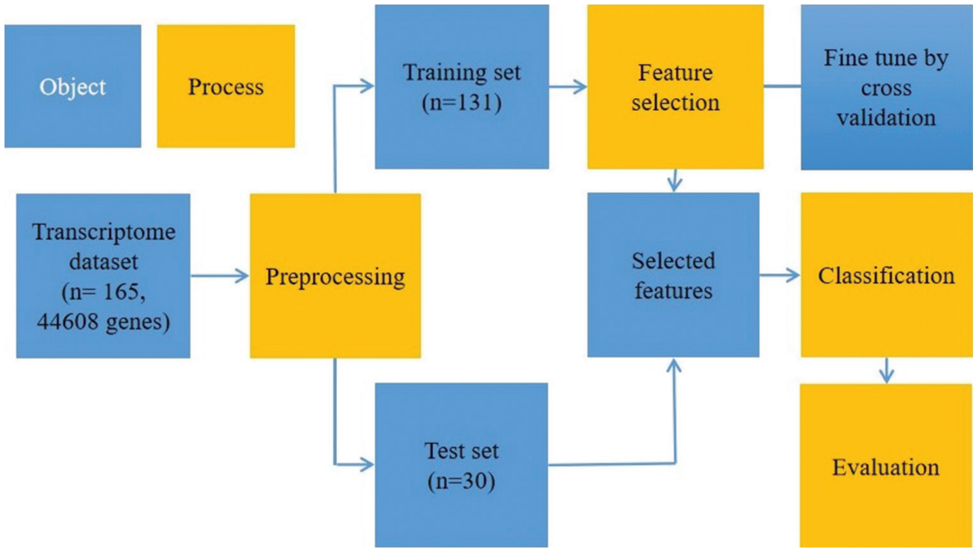


Figure 3. A comparison of two atopic dermatitis classification pipelines by Jiang *et al.*[50] (only the transcriptome dataset)

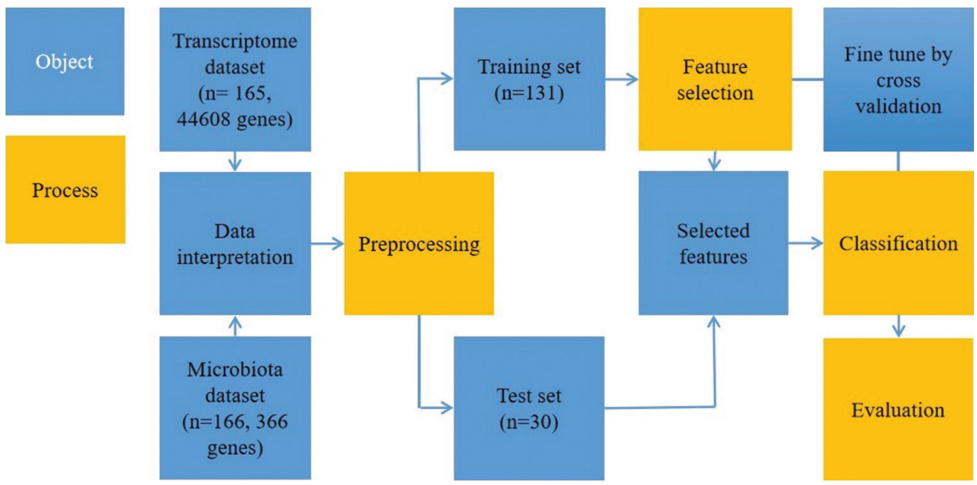


Figure 4. A comparison of two atopic dermatitis classification pipelines by Jiang *et al.*[50] (Both transcriptome and microbiota data)

specificity (85%) leads to a notable and unsatisfactory false positive rate of 15%. Furthermore, it is noteworthy that the ANN model shows diminishing resilience, and the model over fitting that is likely responsible for the provided accuracy is demonstrated by the F1 score of 0.964 and the Matthews correlation coefficient of 0.7454. A thorough summary of the performance results from the further validation of ANN is provided in Table 1, which also sheds light on the model's advantages and disadvantages.

AI-based techniques for diagnosing AD face various challenges stemming from diverse methodologies and data types. Although MPT coupled with AI has demonstrated success in automated diagnosis, it faces challenges such as robust generalization, real-time application feasibility, and ethical concerns. Similarly, serum biomarkers and multivariate ML techniques encounter issues, including selection and integration challenges, data standardization, and the necessity for extensive clinical validation. The complexity of AD, which is a multifactorial condition, renders accurate diagnosis and disease severity evaluation difficult using a single biomarker. Furthermore, integrating gut microbiota, host gene expression, and ML presents challenges concerning data reproducibility and validation across different populations. Replication studies are crucial for validating genes and microbiota traits as viable biomarkers. For instance, the ANN developed by Dautović *et al.*[51] exhibits sensitivity but grapples with specificity and resilience issues. Concerns such as overfitting and model validation necessitate thorough evaluation and validation across diverse datasets. Common challenges persist across these techniques, including the acquisition of diverse,

Table 1. Artificial neural network system performance

	Predicted positive output	Predicted negative output	Output class
Actual positive: 160	True positive (TP):153	False negative (FN): 7	1 – Subjects with disease Output class 0 – Healthy subjects
Actual negative: 20	False positive (FP):3	True negative (TN): 17	Accuracy: 94.44%
Σ = 180	Sensitivity: 95.62%	Specificity: 85%	F1 score: 0.9684
			MCC: 0.7454

Abbreviation: MCC: Matthews correlation coefficient.

representative datasets, ensuring the interpretability and explainability of AI models, addressing ethical considerations, and conducting rigorous clinical validation. The fulfillment of these requirements is essential to ensure the reliability and generalizability of proposed diagnostic tools for AD.

5. AI for monitoring AD

Clinical professionals' subjective visual inspections are frequently used to determine the severity of AD, which introduces significant inter- and intra-observer variability, especially in varied clinical study settings. In an attempt to standardize and automate the diagnosis of AD severity, Pan *et al.*[52] presented EczemaNet, a CNN computer vision pipeline. EczemaNet operates by initially identifying areas affected by AD in images, and subsequently generating probabilistic predictions regarding the severity of the condition. To generate its final predictions, EczemaNet uses ensemble approaches, including crops, ordinal categorization, and transfer and multitask learning. During evaluation in a published clinical trial, EczemaNet exhibited minimal root mean square error and well-calibrated prediction intervals. The research demonstrated the effectiveness of CNNs in treating non-neoplastic skin conditions, especially when dealing with medium-sized datasets. This finding highlights their potential for delivering an objective and more effective assessment of AD severity, which is a development with greater clinical significance compared to simple classification techniques.

Padilla *et al.*[53] used the MobileNet architecture, a CNN, to successfully distinguish between psoriasis and AD. They utilized publicly accessible dermatology to train the network. In a real-world experiment involving a Raspberry Pi camera and 30 subjects, the model successfully classified psoriasis with an impressive 90% accuracy rate and AD with an 88% accuracy rate. In a related study, Patella *et al.*[54] employed an ANN to examine the association between the severity of AD and exposure to air pollutants and environmental factors. Their results revealed a robust association, with the severity of AD lesions increasing by a considerable 200% in response to an increase in the diurnal temperature range, defined as the difference between the highest and lowest temperatures of the day. By predicting disease severity based on environmental parameters, the ANN exhibited promise in providing patients with early warnings to avoid potential irritants, in line with the overarching objective of predictive model-informed, tailored health-care actions.

Neural network algorithms provide a reliable and non-invasive method for classifying AD, frequently demonstrating efficacy in assessing lesion severity. This capability eliminates the need for direct clinical or specialized dermatologist intervention by allowing individuals to remotely monitor their condition using the cameras on their mobile phones. However, it is imperative to acknowledge that ANNs are not designed for constant observation, and existing models generally attain a maximum documented accuracy of 90%. Therefore, future research should prioritize enhancing accuracy and developing user-friendly mobile devices or applications to enable patients to confidently assess their condition using cutting-edge algorithms. In addition, additional scientific evidence is required to determine how the use of emollients or moisturizing creams affects the sensitivity of ANN. These studies could yield important information regarding whether these neural network systems can be used to measure the effectiveness of dermatitis treatment progress. Table 2 provides an overview of key information from the included studies, while Figure 5 illustrates the various data formats used in the previously described studies.

6. Implications of AI in evaluating pruritus in AD

The sensation of itching, medically referred to as pruritus, is a complex problem that significantly affects one's overall quality of life. The persistent and intense prickly discomfort associated with AD can cause mental health issues. These problems may manifest as increased activity levels, generalized anxiety, and, in certain cases, major depressive disorders.[55-57] Despite the profound detrimental effects of pruritus in AD, the lack of standardized and established techniques for objectively evaluating it poses a challenge to physicians and researchers.[58] Pruritus is inherently subjective, with diagnosis primarily reliant on patient reporting. To address this issue, numerous metrics and surveys have been developed. One such commonly used tool is the Peak Pruritus Numerical Rating Scale (NRS),

Table 2. Overview of studies investigating diagnosis and monitoring of AD utilizing AI

Serial No.	Applications	Description	References
1	CRM-PLS-DA	Machine learning-assisted CRM along with PLS-DA aid in precise dermatological diagnosis of AD, distinguishing AD from healthy individuals and exploring multiclass categorization of eczema severity.	33
2	CNN-based MPT	An advanced learning system has been developed to diagnose AD using MPT images, eliminating the need for manual intervention.	34
3	Multivariate machine learning for AD severity prediction using combined biomarkers	The study utilized multivariate machine learning to create a diagnostic tool and severity prediction model for AD patients, revealing intricate connections between clinical and serum measures and enhancing disease understanding.	42
4.	Machine learning classifier for AD detection using gut microbiota and transcriptome data	The study uses a machine learning classifier to accurately detect AD using gut epithelial colonocytes and gut microbiota data. The robust pipeline includes techniques like feature selection, model selection, cross-validation, classification, and statistical assessments, enabling precise discrimination based on omics data.	50
5	ANN for automated AD diagnosis	The ANN has been developed for automated diagnosis of AD. It uses a feed-forward architecture with nine input and output parameters, aiming to improve accuracy and efficiency by distinguishing AD from other skin disorders and using its unique features for classification purposes.	51
6	EczemaNet	EczemaNet is a computer vision tool used to monitor AD severity. It uses CNN and transfer learning techniques to predict severity. Clinical trials confirm its efficacy, providing a standardized, effective monitoring approach.	52
7	Itch Tracker	The Itch Tracker is a software application for smartwatches that tracks nocturnal scratching and provides an objective assessment of itching. It uses an algorithm to analyze acceleration data, distinguishing scratching from other movements based on unique wrist motions. The device is effective in measuring pruritus severity in patients with AD.	71
8	Neurological imaging	The application uses neurological imaging, specifically positron emission tomography and functional magnetic resonance imaging, to objectively identify structural and functional changes in pruritus, detecting brain activity during itching episodes.AI techniques could be used in the analysis of neurological imaging data for pruritus research. These could include automated image analysis, pattern recognition, and predictive modeling.	75
9	Acoustic surveillance	Acoustic monitoring has been used to analyze scratching behavior in AD patients. Initially applied to transgenic mice, a software application automates data analysis. A sound detector integrated with wrist monitoring has accelerated data analysis, but more research is needed.	67,68

Abbreviations: AD: Atopic dermatitis; AI: Artificial intelligence; CNN: Convolutional neural network; MPT: Multiphoton tomography; CRM: Confocal Raman microspectroscopy; PLS-DA: Partial least squares discriminant analysis; ANN: Artificial neural network.

which asks patients a single question: "On a scale of 0 – 10, with 0 being 'no itch' and 10 being 'worst itch imaginable,' how would you rate your itch at the worst point over the past 24 h?" The peak pruritus NRS has proven to be a well-defined, reliable, sensitive, and accurate scale for determining the intensity of the most severe itching. Its clear and simple format makes it particularly appealing to busy clinicians.[59]

Although scales and questionnaires have proven quite useful in clinical settings, biases resulting from individual differences in perceiving and expressing pruritus diminish their utility in research contexts. For example, what one individual may rate as a "10 out of 10" or the "worst itch conceivable" might be noticeably less severe for another, rendering such measures subjective. In addition, not everyone experiences pruritus solely as an "itchy" sensation; some describe it as a "burning" or "tingling," which could cause some clinical instruments to overlook genuine cases of pruritus. These cases demonstrate the limitations of relying solely on scales and questionnaires. While they may aid a doctor in managing an individual AD patient, more objective and supplementary tools are desperately needed.[60-64] More specifically, rigorous research in this area requires instruments capable of identifying minute differences and facilitating comparisons among cohorts before and after AD therapies or interventions.[65,66]

7. Acoustic surveillance

Initially, acoustic surveillance was used on transgenic mice specially engineered to mimic AD. The scratching behavior of these animals was recorded using a sound recording device, and an in-depth analysis of the recorded scratching noises, including the examination of frequency and wavelength data, was conducted. Subsequently, a software application was developed to identify and measure the scratching habits of the mice. This method offers an

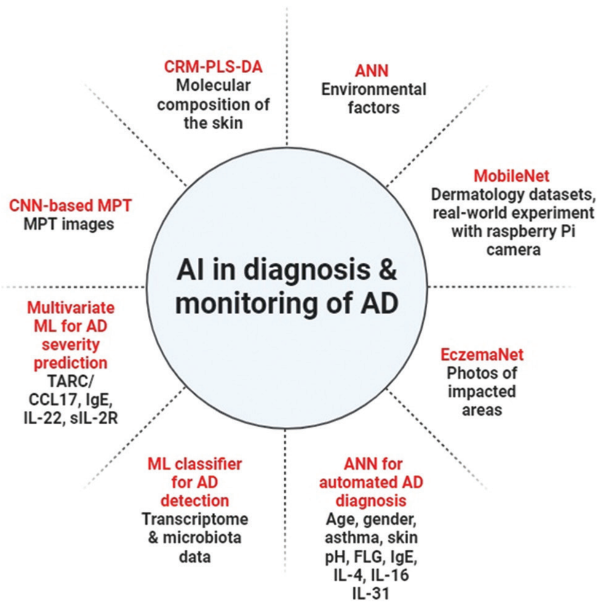


Figure 5. Diverse data modalities (such as images, genetic data, and biomarkers) employed across distinct AI studies.
Abbreviations: AD: Atopic dermatitis; AI: Artificial intelligence; ANN: Artificial neural network; ML: Machine learning; MPT: Multiphoton tomography.

automated and significantly more rapid analysis of scratch data, akin to video surveillance but without the need for direct observation. By contrasting the results of the acoustic counting method with those from video surveillance in the mouse model, the efficacy of the method was confirmed, revealing a close similarity between the two methods.[67]

In 2014, Noro *et al.*[68] expanded on the achievements of the mouse model by developing a sound detector integrated with a similar acoustic surveillance system, designed to be worn on the wrist to monitor scratching behavior. This groundbreaking technology distinguished a particular bone-conducted sound produced by motions through the finger and wrist bones, rather than relying on air-conducted noises. Over a 6-h sleep period, sound data were gathered from both AD patients and healthy controls, while infrared video captured individuals' scratching motions. The algorithm swiftly analyzed the audio data, taking only a few minutes compared to the several hours required for human observers to score video recordings. According to the study, the scratching time recorded by the sound detector was later discovered to be almost identical to the outcomes from video surveillance, which is considered the gold standard for comparison. As a result, the sound detector greatly accelerated and improved the objectivity of the study while also reducing the invasive need for human behavioral observation. Similar to advances in video surveillance, machine-learning algorithms are being actively researched to improve acoustic monitoring. These techniques analyze sound data to detect movement and quantify scratching activity. Despite the advantages of acoustic recording over visual surveillance, it has only been tested in a restricted capacity in the brief trial described earlier. Therefore, more extensive and comprehensive research is needed to confirm this strategy across a wider range of patient demographics.[68]

8. Smart devices

In the past 10 years, medical research has gained insights into a range of medical conditions, and novel therapies have been introduced subsequently, which include monitoring dietary intake and sending reminders for medication adherence, facilitated by the portability and processing power of smart devices.[69,70] Researchers investigating pruritus have modified wrist actigraphy for smartwatches while maintaining its core principles and keeping up with technological advancements. In a pilot study conducted by Lee *et al.*,[71] an accelerometer-equipped wristwatch was used with three subjects to identify scratching tendencies. Remarkably, when compared to infrared video surveillance, the wristwatch demonstrated remarkable accuracy, with detection rates ranging from 98.5 – 99.0% for right-hand scratching motions and 93.3 – 97.6% for left-hand scratching.

In 2017, the Itch Tracker device was developed through a joint effort between dermatologists, Nestle Skin Health, and Apple Inc. This invention consists of an application (software program) designed for well-known smartwatches, aimed at tracking nocturnal scratching and addressing the need for improved objective techniques in assessing itching. As stated in the study, the application incorporates an algorithm that evaluates acceleration data from smartwatches, using unique wrist motions to distinguish scratching from other types of movement. In addition, the application features a smartphone interface that enables users to respond to surveys linked to itching, thus combining subjective patient comments with objective data from wearables.[71] These results highlight the effectiveness of Itch Tracker as a smart gadget for tracking scratching, offering an objective and indirect measure of pruritus severity in patients with AD. With its subtle design and user-friendly interface, Itch Tracker is appropriate for general use. However, further studies are necessary to evaluate the application in various clinical settings and patient demographics.[72]

9. Neurological imaging

A unique method for objectively identifying structural and functional changes in both acute and chronic pruritus is neurological imaging. Positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) have been used to measure brain activity during experimentally induced itching episodes, offering insights into treatments for acute itching.[73] In a previous study, eight patients with AD and six healthy controls underwent PET scanning to examine acute histamine-induced itching. PET scans of AD patients revealed increased brain activity, especially in the basal ganglia, which are known to be a key regulator of the itch-scratch cycle.[74]

A study using arterial spin labeling fMRI, which involved histamine induction in seven healthy controls and eight AD patients, has yielded similar findings, demonstrating markedly enhanced cerebral perfusion after acute scratching in patients with AD.[75] In addition, two other studies investigating brain imaging in individuals with chronic scratching revealed increased activity in reward circuits and motor-related brain areas during scratching.[76,77] Moreover, patients with chronic itching exhibited reduced sensation of itch in response to scratching in these two studies that used arterial spin labeling fMRI to measure brain activity during active scratching. Concurrently, brain regions linked to the central reward system showed significant activation. Taken as a whole, these results emphasize the pleasurable aspect of scratching in relieving both acute and chronic pruritus, and they suggest the possibility that individuals with AD may develop an addiction to scratching.

It is important to recognize three main limitations of neurological imaging in the context of AD-related acute and chronic pruritus.[73] First, many of these studies involve limited sample sizes, and the results are contingent on the techniques used to induce itching and perform imaging. Second, there is a scarcity of studies demonstrating differences in resting states between healthy controls and AD patients. Instead, these techniques need an itch stimulus other than AD-induced itch (e.g., histamine-induced itch) to identify changes that might not accurately mimic the natural state. Third, tests such as PET and fMRI are expensive for the health-care system and uncomfortable for patients, requiring extensive time for execution and analysis. These variables diminish the usefulness of brain imaging for routine clinical diagnosis and standard care. The different distribution of skin lesions with varying morphologies, intensities, and durations might complicate the diagnosis of AD further. However, in recent years, there has been a rapid development in AI-based techniques for image analysis. One method that represents complicated patterns is the use of CNNs, which rely on DL algorithms to identify correlations between neighboring images and integrate them into successive layers. CNNs are used in image analysis across a variety of medical specialties, including cardiology, neurology, and gastrointestinal research, in addition to dermatology.[34]

10. Limitations of AI in dermatology and possible solutions

At present, several noteworthy obstacles hinder the effective application of AI in the medical field, especially in dermatology. One of the primary challenges is the lack of appropriate quality manual annotation and limited sample numbers in existing dermatology training datasets for AI algorithms. This deficiency diminishes the accuracy and usefulness of AI algorithms, rendering them inadequate to meet the demands of routine clinical applications.[78] Furthermore, AI algorithms, usually developed using pre-existing samples, often fail to align with actual medical needs, resulting in a disconnect between them and practical clinical requirements. High-quality training sets are necessary for AI to learn from experience and evolve over time-a capability that human doctors possess naturally. Without such datasets, AI's capacity is hampered, making it challenging to satisfy the growing expectations of both clinical and scientific domains, especially in areas such as the hairy scalp, mucosal membranes, uncommon skin disorders, and the identification of picture artifacts such as colorful marks and tattoos on the skin.[79] Furthermore, training AI to recognize and diagnose a variety of skin problems is difficult due to the wide range of dermatological diseases and the lack of standard criteria for identification and diagnosis.[80] There remains a bottleneck in using AI for the automatic recognition and diagnosis of various dermatopathological images, with current AI applications more frequently used for differentiating between normal and abnormal instances.[81,82] In addition, another major barrier to the use of AI in dermatology is the presence of rare disorders, characterized by a low number of cases and insufficient specimens for appropriate ML training.[83]

At the center of these challenges, it lies the critical importance of guaranteeing the quality of data used in AI services. This intricacy is compounded by intertwined issues, such as potential inaccuracies in the accuracy and reliability of annotations, which can affect the accuracy of the model training assumptions. Furthermore, variability introduced in data collection procedures aggravates these problems, resulting in datasets that are highly inconsistent and pose significant challenges for applying general models to a multitude of clinical scenarios. The presence of such errors, artifacts, or the lack of proper data preprocessing algorithms adds another layer of complexity to raw dermatological data, which must be addressed through comprehensive data preprocessing strategies to enhance the reliability of the model outputs.[84]

One major challenge is addressing bias and limitations in dataset representation, which are evident in the diversity in skin types, conditions, and demographic factors. These aspects may not have been adequately emphasized in training datasets, leading to biases in model outcomes.[85] The complexity adds an extra layer to the dynamics related to clinical practices, impacting data relevance over time and asking for continuous adaptation of AI models. In addition, providing access to various datasets is crucial for adequate model training, but it poses challenges, and potential limitations may restrict effective model generalization. Moreover, ethical considerations and patient privacy issues further complicate matters, especially concerning sensitive dermatological information. Careful balancing of the use of patient information for AI research with individual privacy protection is essential. These factors make it challenging to predict the performance of AI models in actual clinical environments, which can significantly differ from controlled research environments due to diverse patient populations, variations in clinical workflows, and the dynamic nature of healthcare. Effectively recognizing and addressing these nuanced limitations in AI applications in dermatology are pivotal for developing models that not only demonstrate technical proficiency but also seamlessly align with the complex yet ever-changing realities of clinical practice.[86]

Overcoming the challenges of small datasets in AI applications in dermatology requires an appropriate and well-designed strategy. Data augmentation is the key approach that enhances dataset quality by transforming images into different forms, thus improving data diversity and allowing the model to learn more effective features.[87] Transfer learning proves advantageous when models initially trained on huge datasets are fine-tuned using particular dermatology datasets, allowing them to gain general knowledge from a wider context.[88] In addition, the incorporation of carefully developed artificial data, resembling the characteristics of dermatological conditions, serves as an effective pathway toward diversification, underscoring the importance of adequate representation.

Active learning adds a new iterative retraining paradigm where the model selectively prioritizes informative or challenging samples during every retraining cycle, thereby refining its performance.[89] Ensemble models, which employ different architectures and hyperparameters, help minimize the effect of limited data by mixing predictions. Collaboration with other institutions, clinics, or research groups, along with data pooling, ensures the development of a broader and more diverse dataset while adhering to established rules on privacy and ethics. Active data collection remains crucial, requiring regular acquisition through collaborations or utilizing telemedicine platforms to expand the spectrum of the dataset.

To reconcile the disparities between source and target datasets in AI scenarios for dermatology, domain adaptation techniques have been applied.[90] These techniques aim to align the distributions of data and increase adaptability without leading to overfitting incapacitation. By focusing on key features and applying an expert-driven targeted approach within the limitations posed by limited datasets, effective solutions can be developed to address limited data.[91] Overfitting can be mitigated by incorporating regularization techniques such as dropout or weight decay during training.[92] Ensuring the high quality of the limited dataset is crucial as it significantly influences the proper exploitation and effective performance of AI models. By refining these strategies through the iterative process that is in line with evolving research and clinical needs, practitioners could effectively overcome the hurdles imposed by limited datasets in AI applications to dermatology.

11. The acceptance of AI in dermatology: Attitude attribute

The application of AI to medical image recognition has garnered substantial attention recently, particularly in the fields of dermatopathology and dermatology. The growing advancements in AI technology make its use as a decision support tool for dermatologists – particularly in diagnosis support – increasingly relevant within the current legal and health-care frameworks. With the growing utilization of AI by both patients and medical professionals, numerous regional and international survey studies have been conducted to gauge perceptions and attitudes. Between January and June 2019, a comprehensive online survey was distributed to 1271 participants across 92 countries. The results revealed that respondents identified dermoscopic images as the most promising application of AI in dermatology. Significantly, 77.3% of participants expressed approval or strong approval of AI's potential in improving dermatology, with 79.8% incorporating AI into their medical education. However, only a minimal 5.5% (70 out of 1,271) agreed or strongly agreed with the notion that AI would replace dermatologists in the near future. A comparable international survey was conducted among dermatopathologists by the same research team, involving 718 respondents from 91 countries. The findings revealed that 84.1% of respondents thought AI should be included in medical education, and 72.3% of respondents agreed or strongly agreed that AI will improve dermatopathology. Only 6.0% of respondents thought AI would eventually replace human pathologists. Interestingly, 79.2% of respondents thought that automated suggestions for diagnosing skin tumors had strong or very strong potential in terms of diagnostic categorization, whereas 42.6% thought that automated detection of mitosis had the highest potential.[93,94]

Patients typically know little about AI than medical professionals. In a qualitative study conducted from May to July 2019, involving 48 patients and semi-structured interviews for analysis, around 60% of participants stated that shorter diagnosis times and easier access to healthcare were the two biggest advantages of AI for skin cancer surveillance. Nonetheless, 40% of participants expressed concerns about potential dangers, including a rise in patient anxiety. The patients identified the major benefits and drawbacks of AI as the ability to deliver more precise diagnoses (33 [69%]) and less precise diagnoses (41 [85%]). Notably, 35 out of 75 patients stated that they would recommend AI to friends and family.[95] In summary, pathologists and dermatologists generally hold an optimistic view of the prospective advantages and effects of AI in the field of dermatology. However, only a minority of respondents within the cohort exhibited a good or exceptional comprehension of AI. While most pathologists expect AI to be most useful in specific tasks rather than offering overall automated diagnostic advice, a majority of dermatologists believe that AI will improve diagnostic capabilities. Overall, only a small percentage (1 – 3%) of pathologists and dermatologists express concern that AI may soon replace them. As long as AI is used in a way that maintains the doctor-patient relationship, patients are amenable to using it to monitor skin conditions.

12. Perspectives and conclusion

The potential of AI in the field of AD presents an opportunity to significantly enhance diagnostic accuracy and provide personalized healthcare. However, several aspects must be addressed before this innovative approach can be seamlessly integrated into routine clinical practice. AI is gaining recognition at a pace in the field of dermatology, with researchers increasingly focusing on developing AI programs that require diverse data sources for training purposes. These data sources include clinical patient data, which encompasses various aspects such as demographics, comorbidities, characteristics of skin lesions, and relevant laboratory and imaging findings. Furthermore, molecular profiles obtained from biopsy data, such as proteomic analysis, provide valuable information.[5,96] Another avenue involves utilizing data from existing literature. Finally, images play a crucial role in the analysis and classification process. Notably, publicly available benchmarking image datasets such as the International Skin Imaging Collaboration and PH2 dermoscopic archives serve as instrumental training resources for AI models.[97-99] These datasets often consist of lesions that have been confirmed by pathology, follow-up examinations, expert consensus, or *in vivo* confocal microscopy, which enhances their reliability.

Prudent consideration is crucial for researchers involved in the development of AI programs, especially when facing challenges related to training datasets. Estimating the optimal number of training images can be challenging, as having an insufficient dataset may compromise the quality of the program, while an excessively large dataset runs the risk of overfitting the ML classifier to the data, limiting its applicability to external datasets. It is important to note that advanced mathematical techniques are available to address these challenges, such as dropout, data augmentation, batch normalization, and others.[92,100,101] These methods play a key role in preventing overfitting and ensuring the robustness and generalizability of the AI program, which holds significant clinical relevance from a scientific perspective. Efficient utilization of the dataset is crucial in achieving the desired accuracy for specific classifications. In addition, the dataset should include a diverse range of images from various demographics to ensure that resulting algorithms have external validity.[102,103] When acquiring images, it is important to consider potential systematic errors such as variations in lighting, tools, or processes, particularly in different clinical settings, to maintain the research's validity beyond its original context. Simplifying the program's classifications to those with significant prognostic implications can help reduce the size of the dataset and the complexity of algorithms.[102,104]

Randomized clinical trials must be carried out to evaluate the potential of new computer methods and DL in large-scale investigations. Given the limited research in this area, these studies are especially important for gathering data on therapeutic benefits and assisting with causal inference.[105] Moreover, addressing unmet demands such as cost-effectiveness and safety concerns is critical before transferring AI technology from research to clinical settings. Robust regulatory procedures are required to guarantee the safe handling and preservation of private information. Another important challenge is ensuring AI-based healthcare is equitable and inclusive. Healthcare AI should be trained and validated using population-representative data to achieve generalizable performance levels.[106] It is crucial to take into account social and health inequalities that can exclude kids from particular groups who typically have limited access to care. Relying mostly on data from majority ethnic groups or patients with high socioeconomic status could introduce bias into AI performance, as the system may pick up diagnostic tendencies from these over-represented groups.[107]

Developing instructional programs to help doctors successfully use and interpret AI products and services through continuous education is another important factor to take into account. Although AI holds great potential to improve diagnosis, there are several possible drawbacks, including the risk of false diagnoses and erroneous risk assessments. Establishing formal collaboration between industry, the research community, and health-care systems is critical. Such collaboration is necessary to handle every important facet of AI and guarantee a smooth transfer from academic research to practical implementation.

In the near future, physicians and AI-guided computers will probably work closely together. AI has the ability to change clinical management by assisting physicians in analyzing patient data on an individual basis, identifying trends in different diagnostic test results, and more.[108] In this setting, the role of medical experts in interdisciplinary teams becomes essential. The implementation of AI algorithms requires the collection and analysis of large amounts of data, which calls for interpretation within a particular healthcare setting. The domain of "augmented intelligence," where people and machines cooperate to enhance the diagnostic workflow, share judgments, and work in concert, is where clinicians and computers can effectively collaborate. This collaboration has the potential to effectively minimize emotional and economic burdens, leading to more efficient patient care in the health sector.

Acknowledgments

Pratheek Jain and Vinay Alva express gratitude towards Pavara Institute of Medical Sciences (PIMS), Loni, Maharastra, for research support. All authors thank the Management of Alva's Education Foundation (AEF), Moodubidire, ATMA Research Centre, Vidyagiri, and BLDE University, Karnataka, for the constant research motivation.

Funding

None.

Conflict of interest

The authors declare that they have no conflicts of interest.

Author contributions

Conceptualization: Farhan Zameer, Raghavendra Hallur Laxmanashetty
Data curation: Ravish Huchegowda, Ali Jawad Akki
Writing – original draft: Pratheek Jain, Kounaina Khan, Vinay Alva
Writing – review & editing: Raghu Anjanapura Venkataramanaiah, Muthuchelian Krishnasamy, Dilip Apturkar

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable.

References

1. Huchegowda R, Huchegowda S, Pramer J, *et al*. Review on artificial intelligence and applications in healthcare. *Neuro Open J*. 2019;6(1):e1-e4.

   doi: 10.17140/noj-6-e010

2. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, *et al*. Artificial intelligence in cardiology: Present and future. *Mayo Clin Proc*. 2020;95(5):1015-1039.

   doi: 10.1016/j.mayocp.2020.01.038

3. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-38.

   doi: 10.1038/s41591-021-01614-0

4. Cichocki A. Chief Editor's foreword to the inaugural issue of artificial intelligence in health. *AIH*. 2024;1(1):2463.

   doi: 10.36922/aih.2463

5. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.

   doi: 10.1038/nature21056

6. DaVeiga SP. Epidemiology of atopic dermatitis: A review. *Allergy Asthma Proc*. 2012;33(3):227-234.

   doi: 10.2500/aap.2012.33.3569

7. Williams H, Stewart A, Von Mutius E, Cookson W, Anderson HR, International Study of Asthma and Allergies in Childhood (ISAAC) Phase One and Three Study Groups. Is eczema really on the increase worldwide? *J Allergy Clin Immunol*. 2008;121(4):947-954.e15.

   doi: 10.1016/j.jaci.2007.11.004

8. Furue M, Chiba T, Takeuchi S. Current status of atopic dermatitis in Japan. *Asia Pac Allergy*. 2011;1(2):64-72.

   doi: 10.5415/apallergy.2011.1.2.64

9. Williams HC, Strachan DP. The natural history of childhood eczema: Observations from the British 1958 birth cohort study. *Br J Dermatol*. 1998;139(5):834-839.

   doi: 10.1046/j.1365-2133.1998.02501.x

10. Paternoster L, Savenije OE, Heron J, *et al*. Identification of atopic dermatitis subgroups in children from 2 longitudinal birth cohorts. *J Allergy Clin Immunol*. 2018;141(3):964-971.

doi: 10.1016/j.jaci.2017.09.044

11. Kiebert G, Sorensen SV, Revicki D, *et al*. Atopic dermatitis is associated with a decrement in health-related quality of life. *Int J Dermatol*. 2002;41(3):151-158.

    doi: 10.1046/j.1365-4362.2002.01436.x

12. Furue M, Chiba T, Tsuji G, *et al*. Atopic dermatitis: Immune deviation, barrier dysfunction, IgE autoreactivity and new therapies. *Allergol Int*. 2017;66(3):398-403.

    doi: 10.1016/j.alit.2016.12.002

13. Nakahara T, Fujita H, Arima K, Taguchi Y, Motoyama S, Furue M. Treatment satisfaction in atopic dermatitis relates to patient-reported severity: A cross-sectional study. *Allergy*. 2019;74(6):1179-1181.

    doi: 10.1111/all.13712

14. Furue M, Onozuka D, Takeuchi S, *et al*. Poor adherence to oral and topical medication in 3096 dermatological patients as assessed by the Morisky Medication Adherence Scale-8. *Br J Dermatol*. 2014;172(1):272-275.

    doi: 10.1111/bjd.13377

15. Murota H, Takeuchi S, Sugaya M, *et al*. Characterization of socioeconomic status of Japanese patients with atopic dermatitis showing poor medical adherence and reasons for drug discontinuation. *J Dermatol Sci*. 2015;79(3):279-287.

    doi: 10.1016/j.jdermsci.2015.05.010

16. Takeuchi S, Oba J, Esaki H, Furue, M. Non-corticosteroid adherence and itch severity influence perception of itch in atopic dermatitis. *J Dermatol*. 2017;45(2):158-164.

    doi: 10.1111/1346-8138.14124

17. Furue M, Iida K, Imaji M, Nakahara T. Microbiome analysis of forehead skin in patients with atopic dermatitis and healthy subjects: Implication of *Staphylococcus* and *Corynebacterium*. *J Dermatol*. 2018;45(7):876-877.

    doi: 10.1111/1346-8138.14486

18. Iwamoto K, Moriwaki M, Miyake R, Hide M. *Staphylococcus aureus* in atopic dermatitis: Strain-specific cell wall proteins and skin immunity. *Allergol Int*. 2019;68(3):309-315.

    doi: 10.1016/j.alit.2019.02.006

19. Furue M, Kadono T. "Inflammatory skin march" in atopic dermatitis and psoriasis. *J Inflamm Res*. 2017;66(10):833-842.

    doi: 10.1007/s00011-017-1065-z

20. Emam S, Du AX, Surmanowicz P, Thomsen SF, Greiner R, Gniadecki R. Predicting the long-term outcomes of biologics in patients with psoriasis using machine learning. *Br J Dermatol*. 2020;182(5):1305-1307.

    doi: 10.1111/bjd.18741

21. Pierce EJ, Boytsov NN, Vasey JJ, *et al*. A qualitative analysis of provider notes of atopic dermatitis-related visits using natural language processing methods. *Dermatol Ther* (*Heidelb*). 2021;11(4):1305-1318.

    doi: 10.1007/s13555-021-00553-5

22. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444.

    doi: 10.1038/nature14539

23. Esteva A, Robicquet A, Ramsundar B, *et al*. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29.

    doi: 10.1038/s41591-018-0316-z

24. Wolf T, Debut L, Sanh V, *et al*. Transformers: State-of-the-Art Natural Language Processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*; 2020. p. 38-45.

    doi: 10.18653/v1/2020.emnlp-demos.6

25. Gu J, Wang Z, Kuen J, *et al*. Recent advances in convolutional neural networks. *Pattern Recognit*. 2018;77:354-377.

    doi: 10.1016/j.patcog.2017.10.013

26. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: An introduction. *J Am Med Inform Assoc*. 2011;18(5):544-551.

    doi: 10.1136/amiajnl-2011-000464

27. Introducing ChatGPT. *Released on November by 2022 OpenAI*. Available from: https://chat.openai.com [Last accessed on 2024 Jan 05].

28. Moor M, Huang Q, Wu S, *et al*. Med-Flamingo: A Multimodal Medical Few-Shot Learner. In: *Machine Learning for Health (ML4H)*; 2023. p. 353-367. Available from: https://proceedings.mlr.press/v225/moor23a.html

29. Li C. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In: *Advances in Neural Information Processing Systems*; 2023. p. 36

30. Tu T. *Towards Generalist Biomedical AI*; 2023. Available from: https://arxiv.org/abs/2307.14334 [Last accessed on 2024 Jan 05].

31. Zhu D, Chen J, Shen X, Li X, Elhoseiny M. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304 [Preprint]; 2023.

    doi: 10.48550/arXiv.2304.10592

32. Bommasani R, Hudson DA, Adeli E, *et al.* On the Opportunities and Risks of Foundation Models. arXiv:2108. [Preprint]; 2021.

    doi: 10.48550/arXiv.2108.07258

33. Dev K, Ho CJ, Bi R, *et al*. Machine learning assisted handheld confocal raman micro-spectroscopy for identification of clinically relevant atopic eczema biomarkers. *Sensors (Basel)*. 2022;22(13):4674.

    doi: 10.3390/s22134674

34. Guimarães P, Batista A, Zieger M, Kaatz M, Koenig K. Artificial intelligence in multiphoton tomography: Atopic dermatitis diagnosis. *Sci Rep*. 2020;10(1):7968.

    doi: 10.1038/s41598-020-64937-x

35. Nomura I, Tanaka K, Tomita H, *et al*. Evaluation of the *Staphylococcal* exotoxins and their specific IgE in childhood atopic dermatitis. *J Allergy Clin Immunol*. 1999;104(2):441-446.

    doi: 10.1016/s0091-6749(99)70390-8

36. Orfali RL, Sato MN, Santos VG, *et al*. *Staphylococcal* enterotoxin B induces specific IgG4 and IgE antibody serum levels in atopic dermatitis. *Int J Dermatol*. 2014;54(8):898-904.

    doi: 10.1111/ijd.12533

37. Kakinuma T, Nakamura K, Wakugawa M, *et al*. Thymus and activation-regulated chemokine in atopic dermatitis: Serum thymus and activation-regulated chemokine level is closely related with disease activity. *J Allergy Clin Immunol*. 2001;107(3):535-541.

    doi: 10.1067/mai.2001.113237

38. Hijnen D, De Bruin-Weller M, Oosting B, *et al*. Serum thymus and activation-regulated chemokine (TARC) and cutaneous T cell-attracting chemokine (CTACK) levels in allergic diseases TARC and CTACK are disease-specific markers for atopic dermatitis. *J Allergy Clin Immunol*. 2004;113(2):334-340.

    doi: 10.1016/j.jaci.2003.12.007

39. Thijs JL, Nierkens S, Herath A, *et al*. A panel of biomarkers for disease severity in atopic dermatitis. *Clin Exp Allergy*. 2015;45(3):698-701.

    doi: 10.1111/cea.12486

40. Krishnan A, Woo CW, Chang LJ, *et al*. Somatic and vicarious pain are represented by dissociable multivariate brain patterns. *Elife*. 2016;5:e15166.

    doi: 10.7554/elife.15166

41. Wager TD, Atlas LY, Lindquist MA, Roy M, Woo CW, Kross E. An fMRI-based neurologic signature of physical pain. *N Engl J Med*. 2013;368(15):1388-1397.

    doi: 10.1056/nejmoa1204471

42. Lee IS, Yeom M, Kim K, Hahm DH, Kang S, Park HJ. Prediction of disease severity using serum biomarkers in patients with mild-moderate atopic dermatitis: A pilot study. *PLoS One*. 2023;18(11):e0293332.

    doi: 10.1371/journal.pone.0293332

43. Kang S, Kim YK, Yeom M, *et al*. Acupuncture improves symptoms in patients with mild-to-moderate atopic dermatitis: A randomized, sham-controlled preliminary trial. *Complement Ther Med*. 2018;41:90-98.

    doi: 10.1016/j.ctim.2018.08.013

44. Kim J, Kwon SK, Lee IS, *et al.* Effect of acupuncture on gut-brain axis parameters in patients with atopic dermatitis: A study protocol for a randomized, participant-and assessor-blind, sham-controlled trial. *Evid Based Complement Alternat Med*. 2021;2021:5584247.

    doi: 10.1155/2021/5584247

45. Litvak Y, Byndloss MX, Bäumler AJ. Colonocyte metabolism shapes the gut microbiota. *Science*. 2018;362(6418):eaat9076.

    doi: 10.1126/science.aat9076

46. Ghosh D, Bernstein JA, Khurana Hershey GK, Rothenberg ME, Mersha TB. Leveraging multilayered "Omics" data for atopic dermatitis: A road map to precision medicine. *Front. Immunol*. 2018;9:2727.

    doi: 10.3389/fimmu.2018.02727

47. Sacco KA, Milner JD. Gene-environment interactions in primary atopic disorders. *Curr Opin Immunol*. 2019;60:148-155.

    doi: 10.1016/j.coi.2019.06.002

48. Lloyd-Price J, Arze C, Ananthakrishnan AN, *et al*. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*. 2019;569(7758):655-662.

    doi: 10.1038/s41586-019-1237-9

49. Kang MJ, Lee SY, Park YM, *et al*. Interactions between IL-17 variants and *Streptococcus* in the gut contribute to the development of atopic dermatitis in infancy. *Allergy Asthma Immunol Res*. 2021;13(3):404-149.

    doi: 10.4168/aair.2021.13.3.404

50. Jiang Z, Li J, Kong N, *et al*. Accurate diagnosis of atopic dermatitis by combining transcriptome and microbiota data with supervised machine learning. *Sci Rep*. 2022;12:290.

    doi: 10.1038/s41598-021-04373-7

51. Dautović A, Dondras B, Dervisbegovic F, *et al*. Diagnosis of atopic dermatitis using artificial neural network. *IFAC Pap OnLine*. 2022;55(4):51-55.

    doi: 10.1016/j.ifacol.2022.06.008

52. Pan K, Hurault G, Arulkumaran K, Williams HC, Tanaka RJ. EczemaNet: A Detection and Severity Assessment of Atopic Dermatitis. In*: International Workshop on Machine Learning in Medical Imaging*; 2020. p. 220-230.

    doi: 10.1007/978-3-030-59861-7_23

53. Padilla D, Yumang A, Diaz AL, Inlong G. Differentiating Atopic Dermatitis and Psoriasis Chronic Plaque using Convolutional Neural Network MobileNet Architecture. In: *IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*; 2019. p. 1-6.

54. Patella V, Florio G, Palmieri M, *et al*. Atopic dermatitis severity during exposure to air pollutants and weather changes with an artificial neural network (ANN) analysis. *Pediatr Allergy Immunol*. 2020;31(8):938-945.

    doi: 10.1111/pai.13314

55. Slattery MJ, Essex MJ, Paletz EM, *et al*. Depression, anxiety, and dermatologic quality of life in adolescents with atopic dermatitis. *J Allergy Clin Immunol*. 2011;128(3):668-671.

doi: 10.1016/j.jaci.2011.05.003

56. Yaghmaie P, Koudelka CW, Simpson EL. Mental health comorbidity in patients with atopic dermatitis. *J Allergy Clin Immunol*. 2013;131(2):428-433.

doi: 10.1016/j.jaci.2012.10.041

57. Thyssen JP, Hamann CR, Linneberg A, *et al*. Atopic dermatitis is associated with anxiety, depression, and suicidal ideation, but not with psychiatric hospitalization or suicide. *Allergy*. 2017;73(1):214-220.

doi: 10.1111/all.13231

58. Silverberg JI. Practice gaps in pruritus. *Dermatol Clin*. 2016;34(3):257-261.

doi: 10.1016/j.det.2016.02.008

59. Yosipovitch G, Reaney M, Mastey V, *et al*. Peak pruritus numerical rating scale: Psychometric validation and responder definition for assessing itch in moderate-to-severe atopic dermatitis. *Br J Dermatol*. 2019;181(4):761-769.

doi: 10.1111/bjd.17744

60. Ikoma A, Rukwied R, Ständer S, Steinhoff M, Miyachi Y, Schmelz M. Neuronal sensitization for histamine-induced itch in lesional skin of patients with atopic dermatitis. *Arch Dermatol*. 2003;139(11):1455-1458.

doi: 10.1001/archderm.139.11.1455

61. Ikoma A, Steinhoff M, Ständer S, Yosipovitch G, Schmelz M. The neurobiology of itch. *Nat Rev Neurosci*. 2006;7(7):535-547.

doi: 10.1038/nrn1950

62. Paus R, Schmelz M, Biro T, Steinhoff M. Frontiers in pruritus research: Scratching the brain for more effective itch therapy. *J Clin Invest*. 2006;116(5):1174-1186.

doi: 10.1172/jci28553

63. Potenzieri C, Undem BJ. Basic mechanisms of itch. *Clin Exp Allergy*. 2012;42(1):8-19.

doi: 10.1111/j.1365-2222.2011.03791.x

64. Sanders KM, Nattkemper LA, Yosipovitch G. Advances in understanding itching and scratching: a new era of targeted treatments. *F1000Res*. 2016;5:2042.

doi: 10.12688/f1000research.8659.1

65. Weisshaar E, Gieler U, Kupfer J, *et al*. Questionnaires to assess chronic itch: A consensus paper of the special interest group of the international forum on the study of itch. *Acta Derm Venereol*. 2012;92(5):493-496.

doi: 10.2340/00015555-1402

66. Ständer S, Augustin M, Reich A, *et al*. Pruritus assessment in clinical trials: Consensus recommendations from the international forum for the study of itch (IFSI) special interest group scoring itch in clinical trials. *Acta Derm Venereol*. 2013;93(5):509-514.

doi: 10.2340/00015555-1620

67. Umeda K, Noro Y, Murakami T, *et al*. A novel acoustic evaluation system of scratching in mouse dermatitis: Rapid and specific detection of invisibly rapid scratch in an atopic dermatitis model mouse. *Life Sci*. 2006;79(22):2144-2150.

doi: 10.1016/j.lfs.2006.07.010

68. Noro Y, Omoto Y, Umeda K, *et al*. Novel acoustic evaluation system for scratching behavior in itching dermatitis: Rapid and accurate analysis for nocturnal scratching of atopic dermatitis patients. *J Dermatol*. 2014;41(3):233-238.

doi: 10.1111/1346-8138.12405

69. Lu TC, Fu CM, Ma M, Fang CC, Turner A. Healthcare applications of smart watches. *Appl Clin Inform*. 2016;7(3):850-869.

doi: 10.4338/aci-2016-03-r-0042

70. Reeder B, David A. Health at hand: A systematic review of smart watch uses for health and wellness. *J Biomed Inform*. 2016;63:269-276.

doi: 10.1016/j.jbi.2016.09.001

71. Lee J, Cho D, Song S, Kim S, Im E, Kim J. Mobile System Design for Scratch Recognition. In: *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*; 2015. p. 1567-1572.

doi: 10.1145/2702613.2732820

72. Ikoma A, Ebata T, Chantalat L, *et al*. Measurement of nocturnal scratching in patients with pruritus using a smartwatch: Initial clinical studies with the itch tracker app. *Acta Derm Venereol*. 2019;99(3):268-273.

doi: 10.2340/00015555-3105

73. Reich A, Szepietowski JC. Pruritus intensity assessment: Challenge for clinicians. *Expert Rev Dermatol*. 2013;8(3):291-299.

doi: 10.1586/edm.13.2

74. Schneider G, Ständer S, Burgmer M, Driesch G, Heuft G, Weckesser M. Significant differences in central imaging of histamine-induced itch between atopic dermatitis and healthy subjects. *Eur J Pain*. 2008;12(7):834-841.

doi: 10.1016/j.ejpain.2007.12.003

75. Ishiuji Y, Coghill R, Patel T, Oshiro Y, Kraft R, Yosipovitch G. Distinct patterns of brain activity evoked by histamine-induced itch reveal an association with itch intensity and disease severity in atopic dermatitis. *Br J Dermatol*. 2009;161(5):1072-1080.

doi: 10.1111/j.1365-2133.2009.09308.x

76. Papoiu AD, Nattkemper LA, Sanders KM, *et al*. Brain's reward circuits mediate itch relief. A functional MRI study

of active scratching. *PLoS One*. 2013;8(12):e82389.

doi: 10.1371/journal.pone.0082389

77. Mochizuki H, Papoiu AD, Nattkemper LA, *et al*. Scratching induces overactivity in motor-related regions and reward system in chronic itch patients. *J Invest Dermatol*. 2015;135(11):2814-2823.

doi: 10.1038/jid.2015.255

78. Ching T, Himmelstein DS, Beaulieu-Jones BK, *et al*. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;15(141):20170387.

doi: 10.1098/rsif.2017.0387

79. Winkler JK, Fink C, Toberer F, *et al*. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol*. 2019;155(10):1135-1141.

doi: 10.1001/jamadermatol.2019.1735

80. Haw WY, Al-Janabi A, Arents BW, *et al*. Global guidelines in dermatology mapping project (GUIDEMAP): A scoping review of dermatology clinical practice guidelines. *Br J Dermatol*. 2021;185(4):736-744.

doi: 10.1111/bjd.20428

81. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology-new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol*. 2019;16(11):703-715.

doi: 10.1038/s41571-019-0252-y

82. Niazi MK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol*. 2019;20(5):e253-e261.

doi: 10.1016/s1470-2045(19)30154-8

83. Steele L, Velazquez-Pimentel D, Thomas B. Do AI models recognise rare, aggressive skin cancers? An assessment of a direct-to-consumer application in the diagnosis of Merkel cell carcinoma and amelanotic melanoma. *J Eur Acad Dermatol Venereol*. 2021;35(12):e877-e879.

doi: 10.1111/jdv.17517

84. Li Z, Koban KC, Schenck TL, Giunta RE, Li Q, Sun Y. Artificial intelligence in dermatology image analysis: Current developments and future trends. *J Clin Med*. 2022;11(22):6826.

doi: 10.3390/jcm11226826

85. Kleinberg G, Diaz MJ, Batchu S, Lucke-Wold B. Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare. *J Biomed Res*. 2022;3(1):42-47.

doi: 10.46439/biomedres.3.025

86. Omiye JA, Gui H, Daneshjou R, Cai ZR, Muralidharan, V. Principles, applications, and future of artificial intelligence in dermatology. *Front Med* (*Lausanne*). 2023;10:1278232.

doi: 10.3389/fmed.2023.1278232

87. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6(1):60.

doi: 10.1186/s40537-019-0197-0

88. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? *Adv Neural Inf Process Syst*, 2014;27:3320-3328.

89. Settles B. *Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison*; 2009. Available from: http://digital.library.wic.edu/1793/60660 [Last accessed on 2024 Jan 09].

90. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2010;22(10):1345-1359.

doi: 10.1109/tkde.2009.191

91. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;27(8):861-874.

doi: 10.1016/j.patrec.2005.10.010

92. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15(1):1929-1958.

93. Polesie S, Gillstedt M, Kittler H, *et al*. Attitudes towards artificial intelligence within dermatology: An international online survey. *Br J Dermatol*. 2020;183(1):159-161.

doi: 10.1111/bjd.18875

94. Polesie S, McKee PH, Gardner JM, *et al*. Attitudes toward artificial intelligence within dermatopathology: An international online survey. *Front Med* (*Lausanne*). 2020;7:591952.

doi: 10.3389/fmed.2020.591952

95. Nelson CA, Pérez-Chada LM, Creadore A, *et al*. Patient perspectives on the use of artificial intelligence for skin cancer screening: A qualitative study. *JAMA Dermatol*. 2020;156(5):501-512.

doi: 10.1001/jamadermatol.2019.5014

96. Koguchi-Yoshioka H, Watanabe R, Fujisawa Y, *et al*. Skin resident memory T-cell population is not constructed effectively in systemic sclerosis. *Br J Dermatol*. 2019;180(1):219-220.

doi: 10.1111/bjd.17100

97. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data*. 2018;5:180161.

doi: 10.1038/sdata.2018.161

98. ISIC-International Skin Imaging Collaboration; 2018. Available from: https://www.isic-archive.com/#!/topwithheader/widecontenttop/main [Last accessed on 2024 Jan 09].

99. Mendonça T, Ferreira PM, Marques JS, Marcal AR, Rozeira J.

PH2-A Dermoscopic Image Database for Research and Benchmarking. In: *2013 35th Annual International Conference IEEE Engineering Medical Biology Society (EMBC)*. United States: IEEE; 2013. p. 5437-5440.

doi: 10.1109/embc.2013.6610779

100. Ioffe S, Szegedy C. Batch Normalization: Accelerating deep Network Training by Reducing Internal Covariate Shift. In: *International Conference on Machine Learning*; 2015. p. 448-456. Available from: http://arxiv.org/abs/1502.03167 [Last accessed on 2024 Jan 10].

101. Lemley J, Bazrafkan S, Corcoran P. Smart augmentation-learning an optimal data augmentation strategy. *IEEE Access*. 2017;5:5858-5869.

102. Han SS, Lim W, Kim MS, Park I, Park GH, Chang SE. Interpretation of the outputs of a deep learning model trained with a skin cancer dataset. *J Invest Dermatol*. 2018;138(10):2275-2277.

doi: 10.1016/j.jid.2018.05.014

103. Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA. Automated dermatological diagnosis: Hype or reality? *J Invest Dermatol*. 2018;138(10):2277-2279.

doi: 10.1016/j.jid.2018.04.040

104. Hogarty DT, Mackey DA, Hewitt AW. Current state and future prospects of artificial intelligence in ophthalmology: A review. *Clin Exp Ophthalmol*. 2019;47(1):128-139.

doi: 10.1111/ceo.13381

105. Angus DC. Randomized clinical trials of artificial intelligence. *JAMA*. 2020;323(11):1043.

doi: 10.1001/jama.2020.1039

106. Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care. A report from the national academy of medicine. *JAMA*. 2020;323(6):509-510.

doi: 10.1001/jama.2019.21579

107. Bergl PA, Wijesekera TP, Nassery N, Cosby KS. Controversies in diagnosis: Contemporary debates in the diagnostic safety literature. *Diagnosis (Berl)*. 2020;7(1):3-9.

doi: 10.1515/dx-2019-0016

108. Ferrante G, Licari A, Marseglia GL, La Grutta S. Artificial intelligence as an emerging diagnostic approach in paediatric pulmonology. *Respirology*. 2020;25(10):1029-1030.

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

The perspectives of eye care professionals on the integration of artificial intelligence in eye care practices: A systematic review

Obehi Suzan Idogen*

Department of Optometry, Faculty of Life Sciences, University of Benin, Benin, Edo, Nigeria

Abstract

Artificial intelligence (AI) technology has recently been integrated into the health-care industry, including in optometry and ophthalmology. This systematic review assessed the opinions (i.e., perspectives, concerns, and degrees of acceptance) of eye care professionals regarding AI integration into eye care practices. The literature search was conducted using the PubMed and MEDLINE databases. A total of 780 related articles were identified. Among these articles, 304 duplicates were removed, 450 articles were excluded after reviewing the abstract, and 18 articles were excluded after reviewing the full text as these articles were not relevant and/or did not report surveys. The remaining eight included studies were assessed accordingly. Most ophthalmologists and optometrists had a positive perception toward incorporating AI into eye care practices, and these professionals shared that AI would effectively enhance clinical eye care practices. However, certain eye care professionals were concerned about the diagnostic accuracy of AI, the high implementation costs, privacy issues, and the quality of AI-integrated patient care. Several eye care professionals also expressed concerns that AI technology could eventually replace some of their major responsibilities in the practice, suggesting that stakeholders should essentially address these concerns and ensure that AI integration in eye care practices is implemented thoughtfully and ethically to maximize its benefits while preserving the quality of patient care. Nonetheless, this systematic review highlighted the predominantly positive attitude among eye care professionals toward AI integration into eye care practices, warranting further research and collaboration between AI developers and eye care professionals to effectively address the current challenges.

Keywords: Artificial intelligence; Perspective of ophthalmologists; Perspective of optometrists; Eye care

***Corresponding author:**
Obehi Suzan Idogen
(obehi.idogen@lifesci.uniben.edu)

Citation: Idogen OS. The perspectives of eye care professionals on the integration of artificial intelligence in eye care practices: A systematic review. *Artif Intell Health.* 2024;1(2): 66-75. doi: 10.36922/aih.2809

Received: January 23, 2024

Accepted: March 6, 2024

Published Online: April 22, 2024

1. Introduction

Artificial intelligence (AI) technology has become a pivotal aspect of the health-care industry, including optometry and ophthalmology.[1-15] AI technology has been widely used in different eye care practices,[16-19] such as managing and analyzing clinical patient data (e.g., eye images) and addressing staff- and management-related issues.[2,20-32] In addition, AI technology has been used for diagnosing eye disorders, leading to a relatively faster and simpler diagnosis of ocular conditions. As a result, eye care professionals can provide more thorough patient care and generally better practices.[33,34]

However, health-care practitioners and stakeholders often feel apprehensive and uncertain about adopting new and advanced technologies, due to various reported reasons.[35-37] Hence, eye care professionals may have their concerns and perceptions regarding the increasing use of AI technology in eye care practices. The evolving perception of AI in eye care necessitates a thorough evaluation of how the technological advancements of AI can better align with established norms and increase their acceptance within the field.

Herein, this systematic review aims to investigate and analyze the opinions of eye care professionals regarding the integration of AI in eye care practices. Through a comprehensive examination of existing literature, this study offers a detailed understanding of the perspectives, concerns, and knowledge of eye care professionals in adopting AI technology in eye care practices.

2. Methods

2.1. Systematic review approach

I employed the narrative synthesis approach to summarize the findings from multiple studies in this systematic review to provide a thorough explanation of the opinions and perspectives of eye care professionals regarding AI integration in eye care practices. This approach was selected to point out important key findings and patterns in the study's subject area and to offer a comprehensive understanding of the various findings among the included studies.

2.2. Literature search

A literature search was conducted between August and December 2023 using the PubMed interface to access articles in the MEDLINE database. The preference for utilizing only PubMed was due to its accessibility and its extensive collection of records relevant to the research topic. The search focused on studies related to the perception of ophthalmologists and optometrists on AI application in eye care and was not limited to articles published within any particular country, as AI technology had been increasingly applied in eye care practices across various countries. I used the Medical Subject Headings (MeSH) search strategy to identify original survey-based studies conducted among optometrists and ophthalmologists, using the following keywords: "ophthalmologist perception of artificial intelligence," "perception of artificial intelligence in optometry," and "applications of artificial intelligence in eye care."

2.3. Study selection

I established selection criteria for the screening process. The inclusion criteria were as follows: articles published in English; studies published between 2018 and 2023; and survey-based studies that evaluated the perception of ophthalmologists and optometrists on AI integration in eye care. The exclusion criterion was studies involving AI application and implementation in eye care but did not include the perception of optometrists and ophthalmologists toward AI technology.

The literature search yielded a total of 780 articles. From these, 304 duplicate articles were excluded from the study. Subsequently, the abstracts of the remaining 576 articles were screened based on the above selection criteria, resulting in 450 excluded articles. The full text of the remaining 26 articles was further screened, resulting in 18 excluded articles that lacked relevance and/or were not a survey-based study. The remaining eight studies were included and assessed accordingly. Figure 1 illustrates the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) article selection process.

2.4. Data extraction

Relevant information was extracted from the eight included articles, that is, authors' names, year of publication, sample size, survey type, the objective of the study, primary findings, and assessment tools (questionnaire type) used to assess the perceptions of eye care professionals toward the application of AI technology in eye care.

3. Results

Our findings on the included articles are summarized in Table 1.[38-45] Collectively, the eye care professionals displayed varying degrees of optimism toward the application of AI technology in eye care. They recognized the potential advantages of AI technology but had some doubts about its diagnostic precision. Scanzera *et al.*[38] conducted a survey among members of the American Academy of Optometry to determine their opinions, reservations, and readiness to use AI in clinical settings. According to the survey results, 66.8% of optometrists were familiar with AI, suggesting the increasing awareness and interest in AI technology advancements. Some optometrists presented opposing viewpoints on AI, and 25.1% of them were concerned that AI could potentially replace them. Most respondents (72.0%) indicated that AI could improve optometric practices, but 53.0% of respondents had doubts about the technology's diagnostic accuracy. In summation, the opposing responses suggest that AI technology development and implementation in eye care should be further assessed, particularly the accuracy of AI-assisted diagnosis.

Gunasekeran *et al.*[39] comprehensively evaluated the perspectives of ophthalmologists regarding AI in managing

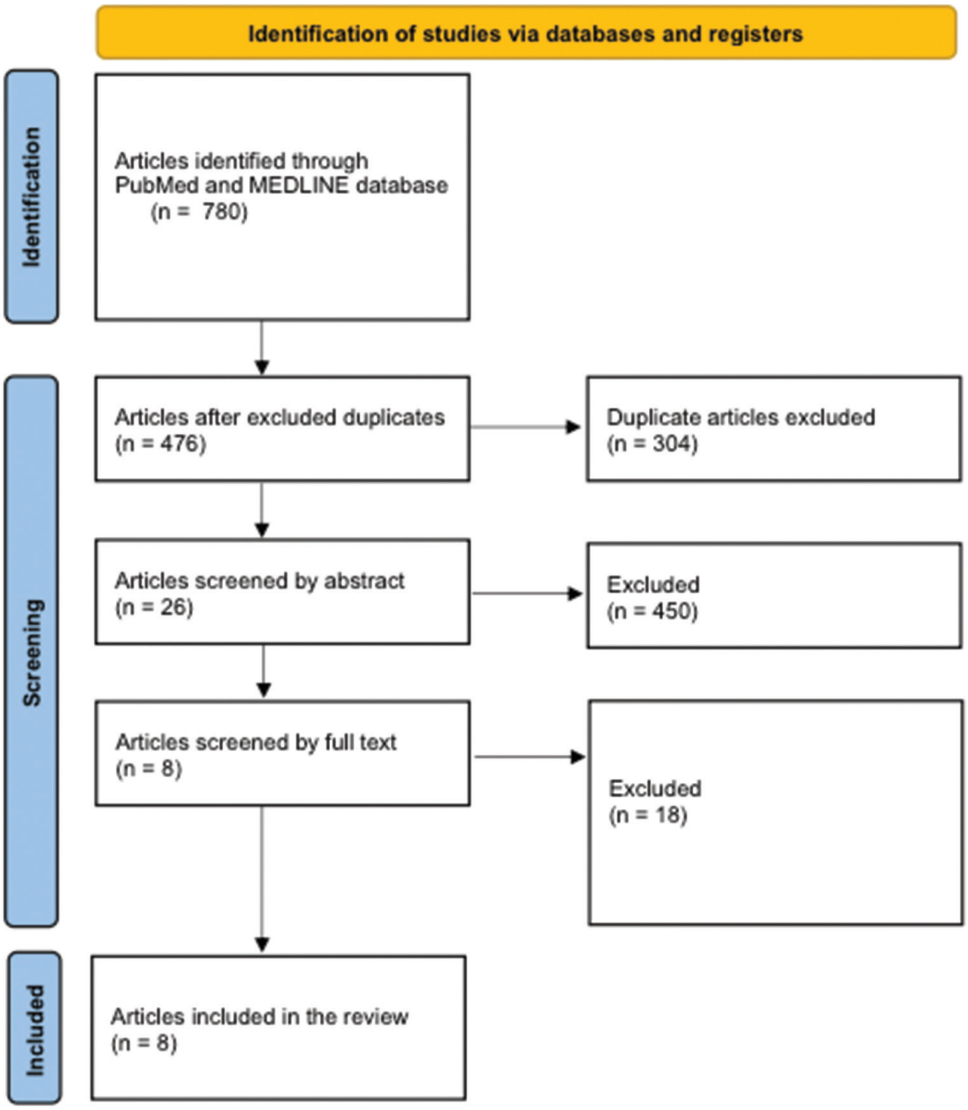


Figure 1. PRISMA diagram illustrating the process of article selection. Adapted from Page *et al*[67]
Abbreviation: PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

prevalent eye conditions, such as diabetic retinopathy, glaucoma, age-related macular degeneration, and cataracts. The global survey involved 1176 ophthalmologists from 70 countries, and the response rates were 78.8–85.8% per question. According to the survey findings, 88.1% of ophthalmologists expressed readiness to use AI technology, particularly as clinical assisting tools. However, the preference for the use of AI as a tool for diagnosis and assisting clinical decisions declined at a response rate of 64.5% and 78.8%, respectively. Most of the respondents expressed confidence that AI would not take their jobs (68.2%). Approximately 72.5% of respondents identified notable challenges in AI implementation, including concerns regarding medical liability resulting from errors. The diagnosis of diabetic retinopathy was identified to be the most significant area for AI application (78.2%), followed by the diagnoses of glaucoma (70.7%), age-related macular degeneration (66.8%), and cataracts (51.4%).

Ho *et al.*[40] assessed the perspectives of optometrists on the use of AI in the diagnosis of retinal disorders. A paper-based survey was conducted among 133 optometrists to determine the factors and obstacles affecting AI implementation in optometry, as well as their general opinion toward AI technology. The primary results of the survey revealed that the surveyed optometrists generally had an optimistic view toward using AI as a support tool to diagnose retinal disorders. The optometrists' perception of AI-assisted diagnosis was positive, with a mean score of 4.0 out of 5 (standard deviation [SD]: 0.8). Furthermore,

Table 1. Summary of the systematic review

Source	Sample size	Objective	Survey type	Findings	Questionnaire
Scanzera *et al.*[38]	400 optometrists	Assess the opinions of optometrists on using AI in eye care	Quantitative	Most optometrists (72.0%) believed that AI could enhance the practice of optometry.	Electronic-based
Gunasekeran *et al.*[39]	1,176 ophthalmologists	Assess the opinions of ophthalmologists on using AI in four primary ocular conditions	Quantitative and qualitative	Ophthalmologists were open to integrating AI as a support tool for managing ocular conditions.	Electronic-based
Ho *et al.*[40]	133 optometrists	Assess the opinions of optometrists on using AI for diagnosing retinal diseases	Quantitative	Optometrists were optimistic about the potential application of AI in diagnosing retinal diseases.	Paper-based
Constantin *et al.*[41]	18 optometrists	Assess the perception of optometrists on utilizing AI in primary eye care and their willingness to provide their patients' ocular scan images to a research repository	Qualitative	Optometrists were willing to share their patients' eye images but raised concerns about technical challenges. They supported AI utilization but emphasized the importance of maintaining their role and responsibilities without diminishment.	Electronic-based
Al-Khaled *et al.*[42]	170 ophthalmologists	Assess the perception of ophthalmologists on AI	Quantitative	Most respondents (75%) reported that AI will advance ophthalmic practice, and 45% of respondents expressed concern about the technology's ability to diagnose patients accurately.	Electronic-based
Scheetz *et al.*[43]	632 healthcare specialists (305 ophthalmologists)	Assess the present utilization, comprehension, and viewpoints on AI in ophthalmology, radiology, and dermatology	Quantitative and qualitative	Most respondents (71%) believed that AI would improve their respective health-care sectors.	Electronic-based
Valikodath *et al.*[44]	80 ophthalmologists	Assess the viewpoints of pediatric ophthalmologists on utilizing AI in ophthalmology	Quantitative and qualitative	Most respondents (70%) were optimistic that AI will enhance ophthalmology practices, while 26% expressed concerns about its potential negative impact on the patient-doctor relationship.	Electronic-based
Alwadani *et al.*[45]	57 ophthalmologists	Assess the knowledge and awareness of junior and senior ophthalmologists on AI	Qualitative	The ophthalmologists had a thorough comprehension of the application of AI in the treatment of eye disorders and had a positive attitude toward it.	Online and paper-based

Abbreviation: AI: Artificial intelligence.

the optometrists were generally agreeable to utilizing AI technology if it had demonstrated improvements in patient access to healthcare (mean score: of 4.4 out of 5). The study also reported that optometrists preferred to use AI to obtain an additional opinion after patient consultation rather than at the point of care. This finding illustrates the potential of integrating AI technology into the patients' therapeutic process to facilitate decision-making with regard to patient care.

Medical images in research archives offer a viable way to improve clinical practice in healthcare.[46] Constantin *et al.*[41] evaluated the attitudes and perspectives of optometrists toward the creation of an image research library and the application of AI to support decision-making in primary eye care. Structured interviews were conducted to determine the expectations, concerns, and recommendations of the eye care professionals with regard to the aim of their study. All of the respondents agreed that retinal images should be contributed to create a comprehensive and long-lasting research archive. The respondents also recognized the possible advantages of the research archive and anticipated improved collaboration between optometry and ophthalmology, particularly with regard to patient referrals. However, the respondents also raised concerns regarding the perceived effort involved in sharing digital photos, lack of standardization, and technical complexity. The respondents displayed their willingness to adopt AI technology to support the diagnosis and treatment of eye disorders. However, the respondents stressed that the application of AI should not diminish their roles and responsibilities and expressed concerns regarding the cost associated with implementing AI technology in eye care.

Taken together, the study demonstrated the willingness among eye care professionals to utilize AI to improve eye care practices but also reflected concerns on the cost of AI implementation, potential job replacements, and the possible compromise in long-standing professional standards.

Several studies have explored the possible effects of AI technology on conventional treatment and doctor-patient interactions.[47,48] For instance, Al-Khaled *et al.*[42] assessed the opinions of 170 ophthalmologists from the United States on incorporating AI into clinical practice. Most of the respondents indicated that they comprehended the prospect of AI technology. However, 22% of respondents were concerned about the effect of AI technology on doctor-patient relationships. The thought that AI would replace doctors was disagreed by 64% of respondents, who also viewed AI as an assistive technology to ophthalmologists instead of their replacement. Approximately 75% of respondents suggested that AI could enhance eye care practices in ophthalmology, and 44% of respondents expressed their interest in using AI in their daily clinical work. However, 45% of ophthalmologists expressed some concerns about the diagnostic accuracy of AI technology, implying their reluctance to depend entirely on AI technology for clinical decision-making.

The attitudes and perspectives of early-career and established clinicians are crucial to the successful integration and adoption of AI in healthcare, as they are key players in its implementation.[49] Scheetz *et al.*[43] conducted a survey with fellows and trainees from three different fields (i.e., ophthalmology, radiology, and dermatology) in Australia and New Zealand to collect their opinions on AI technology. A total of 632 complete responses were obtained, offering valuable insights into the general opinions held by doctors about the potential and influence of AI technology in their respective disciplines. The findings revealed that 71% of respondents expressed optimism that AI would improve medicine, and 85.8% of respondents agreed that AI would directly influence the medical workforce within the next 10 years. AI technology was generally perceived to enhance diagnostics and automate routine clinical procedures, thereby improving the accuracy and efficiency of the clinical sector. However, the respondents also expressed several concerns about AI implementation, such as the possible transfer of health-care duties to tech firms, which could compromise patient privacy, quality of care, and decision-making authority.[43]

Valikodath *et al.*[44] assessed the opinions of 80 pediatric ophthalmologists regarding the use of AI in pediatric ophthalmology. The survey results revealed that the ophthalmologists had a generally optimistic view of AI technology, and 91% of them were familiar with the concept of AI in eye care, demonstrating the foundational knowledge of AI technology developments among the surveyed ophthalmologists. The respondents generally expressed positive views on the possible advantages of AI in ophthalmology. Interestingly, 70% of respondents were confident that AI could improve pediatric ophthalmology practice. Furthermore, 68% of them would be open to integrating AI into their clinical practices, indicating their readiness to use AI technology for improving patient care and diagnosis. Approximately 65% of them did not perceive AI as a replacement for physicians but rather as a complementary tool in clinical decision-making. This perspective aligns with the belief that AI would augment rather than diminish the role of healthcare professionals in delivering quality care to patients. Nonetheless, a minority of the surveyed group expressed some concerns; 15% of respondents expressed concerns regarding the possibility of AI technology replacing medical professionals and negatively impacting the workforce of the healthcare sector;[50,51] 26% of respondents were concerned that AI could undermine the doctor-patient relationship. These concerns highlight the value of preserving human relationships in healthcare.

Alwadani *et al.*[45] conducted a survey among 57 ophthalmologists (i.e., consultants, residents, and fellows) to evaluate their understanding of AI applications in ophthalmology. The questionnaire addressed a wide range of topics, such as demographics, ophthalmology-related expertise, and opinions toward AI. Approximately 91.26% of respondents were confident about the critical role of AI in treating a range of eye conditions. They emphasized the value of AI in the diagnosis and treatment of conditions, such as glaucoma, strabismus, and cataracts. The findings highlighted the respondents' remarkably high knowledge and favorable opinions about the use of AI in treating the aforementioned ocular conditions, implying the significance and advantages of integrating AI education into training programs. Overall, ophthalmologists are becoming increasingly aware of the use of AI in improving patient care, its therapeutic effectiveness, and diagnostic accuracy.

4. Discussion

The integration of AI in eye care practices is gradually becoming more prevalent in certain regions worldwide.[52-54] This review aimed to evaluate the reception of eye care professionals toward embracing AI technology and acknowledging its pivotal role in eye care practices. The utilization of AI technology spans various aspects of eye care, such as the diagnosis of eye diseases, ocular image analysis, and the management of eye care practices.[16,26,32]

Moreover, several studies have highlighted the significant advancements in eye care practices through the integration of AI technology.[17-19,34]

This systematic review revealed that eye care professionals have typically demonstrated favorable enthusiasm and openness toward the integration of AI technology into their clinical practices, considering the remarkable progress AI has made in the field of ophthalmology and optometry.[39,40,44,45] The gradual acceptance of AI technology among eye care professionals suggests that AI technology could potentially establish itself as a fundamental component within eye care practices. Therefore, the opinions of clinicians, as elucidated in this review, are imperative for maximizing the clinical utility of AI and ensuring its successful implementation into routine eye care practice.

Nonetheless, significant concerns were also raised among eye care professionals regarding AI applications. Some professionals indicated the potential inaccuracy and reliability issues of AI technology in diagnosing eye conditions within the clinical setting.[38,42] These concerns raised are valid as precise medical diagnoses are crucial for implementing the right treatment procedures and ensuring patient safety in any medical practice.[55,56] This negative viewpoint sheds light on the diverse perspectives and preferences of eye care professionals toward implementing AI technology in their daily eye care practices, essentially enabling us to identify the gaps in implementing AI in eye care practices.

The cost of implementing AI technology is another significant concern among eye care professionals. Despite their willingness to embrace AI in eye care, the financial limitations in accessing this technology were noted by the eye care professionals.[41] A study investigated the cost of AI implementation and emphasized that the practitioners would have to evaluate whether the benefits would outweigh the cost of AI implementation.[57] This highlights the business aspect of clinical practice, where stakeholders typically strive to balance the cost of production and service delivery with profit.[58]

Another notable concern was the potential replacement of eye care professionals by AI technology.[39,44] A survey among ophthalmologists revealed that most of them were not worried about their roles being replaced, but they did acknowledge the likelihood of certain primary eye care jobs and responsibilities being partially taken over by AI.[17] The study by Valikodath *et al.*[44] indicated that AI is more of an assisting tool rather than a complete replacement. However, AI is increasingly integrated into various aspects of life in the present day, thereby justifying the apprehension of AI technology potentially replacing primary eye care professionals.

Furthermore, health-care professionals, including ophthalmologists, have expressed concerns regarding patient privacy and care quality on AI integration.[43,44] Patient well-being and privacy stand as crucial pillars in healthcare,[59,60] and it is generally believed that human involvement in health-care delivery surpasses AI due to factors such as confidentiality, empathy, adaptability, and experiences only inherent in human health professionals.[61,62] These aspects collectively influence the willingness of health-care practitioners to fully embrace AI in patient care.[63-67]

Nonetheless, the generally positive viewpoints expressed by eye care professionals toward adopting AI highlight the necessity for targeted interventions to effectively address the challenges associated with implementing AI in eye care. To facilitate this, it is crucial to consider initiatives, such as increased funding specifically allocated for AI advancements in eye care. This funding would support research, development, and implementation of AI technologies tailored to the needs of the eye care sector. Moreover, it is important to also emphasize the pivotal role of AI in improving eye care outcomes through education and training programs designed for eye care professionals. By imparting knowledge about AI technology, its benefits, and its potential impact on eye care, professionals can develop a deeper understanding of how to integrate these technologies into their practice. The training programs should also address concerns surrounding the accuracy of AI diagnoses and their effect on patient care. Furthermore, it is beneficial to foster an environment that encourages the adoption of AI technology while actively addressing apprehensions and doubts within the eye care community. This could be achieved by establishing channels for open communication, information exchange, and collaborations between researchers, technologists, and eye care specialists to overcome current limitations and realize the full potential of AI in the field of eye care.

5. Conclusion

This systematic review demonstrated the generally positive perception of most eye care professionals regarding the implementation of AI technology in eye care practices. However, there were several key concerns raised, such as the high costs associated with implementing AI, the uncertain reliability of AI in performing diagnoses and making clinical decisions, and the fear of AI potentially replacing eye care professionals. Overall, this study highlighted the importance of optometrists and ophthalmologists embracing technological advancements such as AI, emphasizing the necessity of addressing the expressed concerns to ensure a harmonious integration of AI technology into eye care.

Acknowledgments

None.

Funding

None.

Conflict of interest

The author declares no competing interests.

Author contributions

This is a single-authored article.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable.

References

1. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthc J*. 2021;8(2):e188-e194.

   doi: 10.7861/fhj.2021-0095

2. Alowais SA, Alghamdi SS, Alsuhebany N, *et al*. Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Med Educ*. 2023;23(1):689.

   doi: 10.1186/s12909-023-04698-z

3. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J*. 2019;6(2):94-98.

   doi: 10.7861/futurehosp.6-2-94

4. Gruson D. L'intelligence artificielle en santé, un potentiel majeur d'innovations pour notre système de santé Artificial intelligence in healthcare: Major potential for innovations in our health system]. *Soins*. 2019;64(838):33-35.

   doi: 10.1016/j.soin.2019.06.006

5. Madison DE. Rapid prototyping for healthcare applications. *Comput Healthc*. 1989;10(11):35-38.

6. Chan Y, Chen Y, Pham T, Chang W, Hsieh MY. Artificial intelligence in medical applications. *J Healthc Eng*. 2018;2018:4827875.

   doi: 10.1155/2018/4827875

7. Basu K, Sinha R, Ong A, Basu T. Artificial intelligence: How is it changing medical sciences and its future? *Indian J Dermatol*. 2020;65(5):365-370.

   doi: 10.4103/ijd.IJD_421_20.

8. Yu YY. Role of artificial intelligence in the diagnosis and treatment of gastrointestinal diseases. *Zhonghua Wei Chang Wai Ke Za Zhi*. 2020;23(1):33-37.

   doi: 10.3760/cma.j.issn.1671-0274.2020.01.006

9. Niel O, Bastard P. Artificial intelligence in nephrology: Core concepts, clinical applications, and perspectives. *Am J Kidney Dis*. 2019;74(6):803-810.

   doi: 10.1053/j.ajkd.2019.05.020

10. Bitkina OV, Park J, Kim HK. Application of artificial intelligence in medical technologies: A systematic review of main trends. *Digit Health*. 2023;9:20552076231189331.

    doi: 10.1177/20552076231189331

11. Yang X, Wu J, Chen X. Application of artificial intelligence to the diagnosis and therapy of nasopharyngeal carcinoma. *J Clin Med*. 2023;12(9):3077.

    doi: 10.3390/jcm12093077

12. Faizal KD, Sultan RF. Applications of artificial intelligence and big data analytics in m-health: A healthcare system perspective. *J Healthc Eng*. 2020;2020:8894694.

    doi: 10.1155/2020/8894694

13. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2(10):719-731.

    doi: 10.1038/s41551-018-0305-z

14. Manickam P, Mariappan SA, Murugesan SM, *et al*. Artificial intelligence (AI) and internet of medical things (IoMT) assisted biomedical systems for intelligent healthcare. *Biosensors* (*Basel*). 2022;12(8):562.

    doi: 10.3390/bios12080562

15. Playford D, Bordin E, Mohamad R, Stewart S, Strange G. Enhanced diagnosis of severe aortic stenosis using artificial intelligence: A proof-of-concept study of 530,871 echocardiograms. *JACC Cardiovasc. Imaging*. 2020;13(4):1087-1090.

    doi: 10.1016/j.jcmg.2019.10.013

16. Lu W, Tong Y, Yu Y, Xing Y, Chen C, Shen Y. Applications of artificial intelligence in ophthalmology: General overview. *J Ophthalmol*. 2018;2018:5278196.

    doi: 10.1155/2018/5278196

17. Tan Z, Scheetz J, He M. Artificial intelligence in ophthalmology: Accuracy, challenges, and clinical application. *Asia Pac J Ophthalmol* (*Phila*). 2019;8(3):197-199.

    doi: 10.22608/APO.2019122

18. Bellemo V, Lim ZW, Lim G, *et al*. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: A clinical validation study. *Lancet Digit Health*. 2019;1(1):e35-e44.

doi: 10.1016/S2589-7500(19)30004-4

19. Shao Y, Jie Y, Liu ZG, *et al*. Guidelines for the application of artificial intelligence in the diagnosis of anterior segment diseases (2023). *Int J Ophthalmol*. 2023;16(9):1373-1385.

    doi: 10.18240/ijo.2023.09.03

20. Krishnan G, Singh S, Pathania M, *et al*. Artificial intelligence in clinical medicine: Catalyzing a sustainable global healthcare paradigm. *Front Artif Intell*. 2023;6:1227091.

    doi: 10.3389/frai.2023.1227091

21. Yoon JH, Pinsky MR, Clermont G. Artificial intelligence in critical care medicine. *Crit Care*. 2022;26(1):75.

    doi: 10.1186/s13054-022-03915-3

22. Jiang F, Jiang Y, Zhi H, *et al*. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230-243.

    doi: 10.1136/svn-2017-000101

23. Currie G, Hawk KE, Rohren E, Vial A, Klein R. Machine learning and deep learning in medical imaging: Intelligent imaging. *J Med Imaging Radiat Sci*. 2019;50(4):477-487.

    doi: 10.1016/j.jmir.2019.09.005

24. Al-Atari MA. Artificial intelligence for medical diagnostics-existing and future AI technology!. *Diagnostics* (*Basel*). 2023;13(4):688.

    doi: 10.3390/diagnostics13040688

25. Miller DD, Brown EW. How cognitive machines can augment medical imaging. *AJR Am J Roentgenol*. 2019;212(1):9-14.

    doi: 10.2214/AJR.18.19914

26. Li H, Cao J, Grzybowski A, Jin K, Lou L, Ye J. Diagnosing systemic disorders with AI algorithms based on ocular images. *Healthcare* (*Basel*). 2023;11(12):1739.

    doi: 10.3390/healthcare11121739

27. Tan Y, Sun X. Ocular images-based artificial intelligence on systemic diseases. *Biomed Eng Online*. 2023;22(1):49.

    doi: 10.1186/s12938-023-01110-1

28. Balyen L, Peto T. Promising artificial intelligence-machine learning-deep learning algorithms in ophthalmology. *Asia Pac J Ophthalmol* (*Phila*). 2019;8(3):264-272.

    doi: 10.22608/APO.2018479

29. Li H, Cao J, You K, Zhang Y, Ye J. Artificial intelligence-assisted management of retinal detachment from ultra-widefield fundus images based on weakly-supervised approach. *Front Med* (*Lausanne*). 2024;11:1326004.

    doi: 10.3389/fmed.2024.1326004

30. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications. *Bioengineering* (*Basel*). 2023;10(12):1435.

    doi: 10.3390/bioengineering10121435

31. Zirar A, Ali IS, Islam M. Worker and workplace artificial intelligence (AI) coexistence: Emerging themes and research agenda. *Technovation*. 2023;124(1):102747.

    doi: 10.1016/j.technovation.2023.102747

32. González-Gonzalo C, Thee EF, Klaver CCW, *et al*. Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice. *Prog Retin Eye Res*. 2022;90:101034.

    doi: 10.1016/j.preteyeres.2021.101034

33. Du XL, Li WB, Hu BJ. Application of artificial intelligence in ophthalmology. *Int J Ophthalmol*. 2018;11(9):1555-1561.

    doi: 10.18240/ijo.2018.09.21

34. Du HQ, Dai Q, Zhang ZH, *et al*. Artificial intelligence-aided diagnosis and treatment in the field of optometry. *Int J Ophthalmol*. 2023;16(9):1406-1416.

    doi: 10.18240/ijo.2023.09.06

35. Peterson L, Larsson I, Nygren JM, *et al*. Challenges to implementing artificial intelligence in healthcare: A qualitative interview study with healthcare leaders in Sweden. *BMC Health Serv Res*. 2022;22(1):850.

    doi: 10.1186/s12913-022-08215-8

36. Cobelli N, Cassia F, Burro R. Factors affecting the choices of adoption/non-adoption of future technologies during coronavirus pandemic. *Technol Forecast Soc Change*. 2021;169:120814.

    doi: 10.1016/j.techfore.2021.120814

37. Jedwab RM, Hutchinson AM, Manias E, *et al*. Nurse motivation, engagement and well-being before an electronic medical record system implementation: A mixed methods study. *Int J Environ Res Public Health*. 2021;18(5):2726.

    doi: 10.3390/ijerph18052726

38. Scanzera AC, Shorter E, Kinnaird C, *et al*. Optometrist's perspectives of artificial intelligence in eye care. *J Optom*. 2022;15 Suppl 1(Suppl 1):S91-S97.

    doi: 10.1016/j.optom.2022.06.006

39. Gunasekeran DV, Zheng F, Lim GYS, *et al*. Acceptance and perception of artificial intelligence usability in eye care (APPRAISE) for ophthalmologists: A multinational perspective. *Front Med*. 2022;9:875242.

    doi: 10.3389/fmed.2022.875242

40. Ho S, Doig GS, Ly A. Attitudes of optometrists towards artificial intelligence for the diagnosis of retinal disease: A cross-sectional mail-out survey. *Ophthalmic Physiol Opt*. 2022;42(6):1170-1179.

    doi: 10.1111/opo.13034

41. Constantin A, Atkinson M, Bernabeu MO, *et al*. Optometrists' perspectives regarding artificial intelligence

aids and contributing retinal images to a repository: Web-based interview study. *JMIR Hum Factors*. 2023;10:e40887.

doi: 10.2196/40887

42. Al-Khaled T, Valikodath N, Cole E, *et al*. Evaluation of physician perspectives of artificial intelligence in ophthalmology: A pilot study. *Invest Ophthalmol Vis Sci*. 2020;61(7):202.
43. Scheetz J, Rothschild P, McGuinness M, *et al.* A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Sci Rep.* 2021;11(1):5193.

    doi: 10.1038/s41598-021-84698-5
44. Valikodath NG, Al-Khaled T, Cole E, *et al*. Evaluation of pediatric ophthalmologists' perspectives of artificial intelligence in ophthalmology. *J AAPOS*. 2021;25(3):164.e1-164.e5.

    doi: 10.1016/j.jaapos.2021.01.011
45. Alwadani AF, Zakaria MO, Alwadany MN, *et al*. Ophthalmologists' view of artificial intelligence: Results of a cross-sectional survey. *Int J Med Dev Countries*. 2023;7(5):811-817.

    doi: 10.24911/IJMDC.51-1673725204
46. Islam MM, Poly TN, Li YJ. Recent advancement of clinical information systems: Opportunities and challenges. *Yearb Med Inform*. 2018;27(1):83-90.

    doi: 10.1055/s-0038-1667075
47. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: Some problems and solutions. *BMC Med Inform Decis Mak*. 2023;23(1):73.

    doi: 10.1186/s12911-023-02162-y
48. Kerasidou A. Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. *Bull World Health Organ*. 2020;98(4):245-250.

    doi: 10.2471/BLT.19.237198
49. Hogg HDJ, Al-Zubaidy M, Talks J, *et al*. Stakeholder perspectives of clinical artificial intelligence implementation: Systematic review of qualitative evidence. *J Med Internet Res.* 2023;25:e39742.

    doi: 10.2196/39742
50. Castagno S, Khalifa M. Perceptions of artificial intelligence among healthcare staff: A qualitative survey study. *Front Artif Intell*. 2020;3:578983.

    doi: 10.3389/frai.2020.578983
51. Orlova IA, Akopyan ZA, Plisyuk A, *et al*. Opinion research among Russian Physicians on the application of technologies using artificial intelligence in the field of medicine and health care. *BMC Health Serv Res.* 2023;23(1):749.

    doi: 10.1186/s12913-023-09493-6
52. Tan TF, Thirunavukarasu AJ, Jin L, *et al*. Artificial intelligence and digital health in global eye health: Opportunities and challenges. *Lancet Glob Health*. 2023;11(9):e1432-e1443.

    doi: 10.1016/S2214-109X(23)00323-6
53. Li JO, Liu H, Ting DSJ, *et al*. Digital technology, tele-medicine and artificial intelligence in ophthalmology: A global perspective. *Prog Retin Eye Res.* 2021;82:100900.

    doi: 10.1016/j.preteyeres.2020.100900
54. Jin K, Ye J. Artificial intelligence and deep learning in ophthalmology: Current status and future perspectives. *Adv Ophthalmol Pract Res*. 2022;2(3):100078.

    doi: 10.1016/j.aopr.2022.100078
55. Balogh EP, Miller BT, Ball JR, editors. *Improving Diagnosis in Health Care*. United States: National Academies Press (US); 2015.
56. Croskerry P. Perspectives on diagnostic failure and patient safety. *Healthc Q*. 2012;15 Spec No:50-56.

    doi: 10.12927/hcq.2012.22841
57. Ruamviboonsuk P, Chantry S, Seresirikachorn K, Ruamviboonsuk V, Sangroongruangsri S. Economic evaluations of artificial intelligence in ophthalmology. *Asia Pac J Ophthalmol (Phila*). 2021;10(3):307-316.

    doi: 10.1097/APO.0000000000000403
58. Gray BH, editors. Institute of Medicine (US). *Committee on Implications of For-Profit Enterprise in Health Care*. United States: National Academies Press (US); 1986.
59. Institute of Medicine (US) Committee on regional health data networks. In: Donaldson MS, Lohr KN, editors. *Health Data in the Information Age: Use, Disclosure, and Privacy*. United States: National Academies Press (US); 1994.
60. Tegegne MD, Melaku MS, Shimie AW, *et al.* Health professionals' knowledge and attitude towards patient confidentiality and associated factors in a resource-limited setting: A cross-sectional study. *BMC Med Ethics*. 2022;23(1):26.

    doi: 10.1186/s12910-022-00765-0
61. Moudatsou M, Stavropoulou A, Philalithis A, Koukouli S. The role of empathy in health and social care professionals. *Healthcare* (*Basel*). 2020;8(1):26.

    doi: 10.3390/healthcare8010026
62. McNulty JP, Politis Y. Empathy, emotional intelligence and interprofessional skills in healthcare education. *J Med Imaging Radiat Sci.* 2023;54(2):238-246.

    doi: 10.1016/j.jmir.2023.02.014
63. Murdoch B. Privacy and artificial intelligence: Challenges for protecting health information in a new era. *BMC Med Ethics*. 2021;22(1):122.

    doi: 10.1186/s12910-021-00687-3

64. Pashkov VM, Harkusha AO, Harkusha YO. Artificial intelligence in medical practice: Regulative issues and perspectives. *Wiad Lek*. 2020;73(12 cz 2):2722-2727.

65. Morrow E, Zidaru T, Ross F, *et al*. Artificial intelligence technologies and compassion in healthcare: A systematic scoping review. *Front Psychol*. 2023;13:971044.

doi: 10.3389/fpsyg.2022.971044

66. Wang C, Zhang J, Lassi N, Zhang X. Privacy protection in using artificial intelligence for healthcare: Chinese regulation in comparative perspective. *Healthcare* (*Basel*). 2022;10(10):1878.

doi: 10.3390/healthcare10101878

67. Page MJ, McKenzie JEM, Bossuyt PM, *et al*. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.

doi: 10.1136/bmj.n71

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Discovery of new antibiotics using AI-guided spectroscopy and 3D drug-protein computer simulation technologies to combat MDR bacteria-associated mortality

Asit Kumar Chakraborty*, Meghna Maity, and Sumana Sahoo

Department of Biochemistry and Biotechnology, Faculty of Oriental Institute of Science and Technology, Vidyasagar University, Midnapore, West Bengal, India

Abstract

Multidrug-resistant (MDR), extensively drug-resistant (XDR), and totally drug-resistant bacteria can cause sepsis and death in patients due to their ability to inactivate most antibiotics, including ampicillin, tetracycline, streptomycin, chloramphenicol, erythromycin, and ciprofloxacin. This paper aims to review recent advancements in synthetic antibiotics, lantibiotics, and phytoantibiotics and to present our research on phytoantibiotics, specifically focusing on CU1 and NU2. While third- and fifth-generation synthetic antibiotics such as meropenem, moxifloxacin, amikacin, and tigecycline are currently relied upon for treating MDR infections, research is underway to develop peptide antibiotics known as lantibiotics (e.g., nisins, bacteriocins, and salivaricins). Lantibiotics such as nisin-A and salivaricin-B have demonstrated efficacy in curing numerous MDR infections, while phytochemicals such as artemisinin and quinine have shown effectiveness against chloroquine-resistant *Plasmodium falciparum* infections (malaria). In our study, we utilized techniques such as mass spectroscopy, nuclear magnetic resonance, and Fourier transform infrared spectroscopy in conjunction with artificial intelligence (AI) and computer simulation technologies to determine the structure of phytochemicals. Our results revealed that CU1, derived from *Cassia fistula* bark ethanol extract, exhibits potent antibiotic activity against XDR bacteria by targeting the RNA polymerases of *Escherichia coli* and *Mycobacterium tuberculosis*. Consequently, our MDR-Cure extract containing CU1 represents a promising antibacterial Ayurvedic medicine specifically tailored for skin and nail infections. Similarly, NU2 poly-fluorophosphate-glycosides from *Suregada multiflora* roots ethanol extract exhibited strong inhibitory effects on XDR bacteria by targeting DNA topoisomerase I. Recently, many cyclic peptide antibiotics have been synthesized *in vitro* using computer-guided AI technologies to predict 3D drug-enzyme interactions and are currently undergoing clinical trials. Our ultimate goal is to combat XDR bacteria-associated deaths, which are predicted to escalate as we approach 2050.

Keywords: Lantibiotics; Phytoantibiotics; Meropenem; Moxifloxacin; Salivaricin; Extensively drug-resistant tuberculosis

***Corresponding author:**
Asit Kumar Chakraborty
(chakraakc@gmail.com)

Citation: Chakraborty AK, Maity M, Sahoo S. Discovery of new antibiotics using AI-guided spectroscopy and 3D drug-protein computer simulation technologies to combat MDR bacteria-associated mortality. *Artif Intell Health.* 2024;1(2): 76-95. doi: 10.36922/aih.2284

Received: November 21, 2023

Accepted: January 17, 2024

Published Online: April 23, 2024

1. Introduction

The penicillin antibiotic was discovered in 1928, leading to the development of numerous derivatives, such as ampicillin, amoxicillin, cefotaxime, imipenem, and meropenem, to combat bacterial infections.[1] Unfortunately, drug-resistance proteins in bacterial extracts were detected as early as 1940. It took 25 years to isolate the *amp* gene in pBR322 and to demonstrate that purified Amp penicillinase could cleave penicillin *in vitro*. Since then, drug companies have faced concerns that their new antibiotics might become obsolete within a few months of their commercial release. Modern multidrug-resistance (MDR) conjugative plasmids in bacteria are associated with mobile elements, integrons, integrases, transposases, and DNA topoisomerases, facilitating the emergence of new *mdr* genes that render newly developed antibiotics ineffective and economically unviable. Despite these challenges, scientists continued to develop new derivatives of penicillin (cephalosporins and carbapenems), aminoglycosides (amikacin), quinolones (moxifloxacin, lomofloxacin), and many others. On the other hand, bacteria have evolved to carry genotypes such as *blaTEM, blaCTX-M, blaOXA-58, blaKPC*, and *blaNDM*, which encode enzymes capable of inactivating all first- to fifth-generation PENEM antibiotics.[2] Similarly, at least 20 TET drug efflux derivatives have been sequenced, and these enzymes efflux tetracycline or its higher derivatives such as doxycycline, minocycline, and tigecycline. Chloramphenicol acetyltransferases (*cat* gene) were first discovered in the inactivation of chloramphenicol by acetylation, followed by the identification of AAC3' and AAC6' enzymes in bacterial plasmids responsible for acetylating streptomycin, amikacin, or erythromycin. The discovery of *StrA* and *StrB,* two linked genes in plasmids, has enhanced our understanding of how these two enzymes phosphorylate streptomycin to inactivate the tuberculosis (TB) drug. Lately, dozens of such isomers (APH2', APH3', and APH6') have been discovered to inactivate most amino-glycoside antibiotics.[3-5]

We also observe conventional point mutations in target enzymes, such as rpoB (RNA polymerase subunit) and gyrA/B (DNA topoisomerase II subunits), which lead to the inhibition of the binding of drugs rifampicin and ciprofloxacin, respectively, to the target enzyme, thereby preventing them from exerting their inhibitory effects. Consequently, the rate of approval for new antibiotic derivatives has decreased significantly, with only 2 – 5 approvals per year compared to hundreds released annually between 1970 and 2000 for commercial therapy of infectious diseases.[6,7]

Lantibiotics are cyclic peptide antibiotics produced by *Streptomyces, Streptococcus, Lactobacillus,* and other bacteria to combat other surrounding bacteria and microbes.[8,9] This diverse class includes gramicidins, salivaricins, mutacins, nisins, bacteriocins, and other lantibiotics, which hold promise for commercial use in combating MDR bacteria. Recombinant technology is now employed to express lantibiotic cyclase, lantibiotic synthetase, and lantibiotic transferase-peptidase to develop special types of cyclic peptide reactions, resulting in the development of novel trypsin-resistant peptide antibiotics against MDR bacteria. To avert MDR, it is imperative to avoid non-prescription and uncontrolled antibiotics and refrain from releasing antibiotics or lantibiotics into rivers, ponds, and seas.[10]

In ancient times, India and China were renowned for their rich tradition of herbal preparations to cure diverse diseases, including bacterial and fungal infections. However, during British rule, India's ancient tradition of herbal drugs waned, while China continued to prioritize herbal drugs.[11] Nevertheless, studies from the United States have demonstrated the efficacy of a phytodrug, artimisinin derivatives, to combat chloroquine-resistant malaria parasites, as well as the ability of phytodrugs, taxol, and topotecan to cure various types of cancer.[12] Despite India's daily publication of numerous papers on phytoextracts with antibacterial activities, commercialization has been hindered by the low bioactive chemical content and poor inhibitory power of such extracts.[13] We made our first important progress in isolating CU1 poly-bromophenol-turpentine from *Cassia fistula* bark targeting RNA polymerase.[14,15] Furthermore, we developed a large-quantity purification process of NU2 poly-fluorophosphate-glycosides from *Suregada multiflora* root using thin-layer chromatography (TLC) and ultraviolet (UV)-shadowing, targeting MDR bacterial DNA topoisomerase I (in preparation). NU2 has also exhibited an inhibitory role against the malaria parasite *Plasmodium falciparum* that resides in the human red blood cell.

Plant secondary metabolites are naturally produced compounds that can inhibit soil bacteria. MDR bacteria are known to proliferate in various environments, including soil, water (pond, river, sea, rain), chicken meat, milk, and human skin and hair. This constant exposure to newer strains of MDR bacteria suggests that plants may continuously produce new antibiotics, making plant extracts an ideal source for developing newer drugs against MDR bacteria and fungus. Our goal is to review the recent development and outcome of synthetic antibiotics, lantibiotics, and phytoantibiotics that are related to our phytodrug development program. However, we have carefully avoided the development of new synthetic antibiotics, as many important reviews are available.[16]

Instead, we emphasized peptide antibiotics, which are novel in their preparation and mode of action.[17]

2. Materials and methods

2.1. PubMed and database search

We conducted a search on PubMed (www.ncbi.nlm.nih.gov/pubmed) using the following terms: "MDR bacteria," "nisin," "salivaricin," "phyto-drugs," "herbal drugs," and "MDR plasmids."

2.2. Isolation of multi-drug-resistant bacteria

We isolated MDR bacteria from samples obtained from the Ganges River by plating 0.1 mL of water onto LB+agar+50 μg/mL ampicillin or other antibiotics. Following incubation, single colonies were selected and tested for various drug sensitivities using different antibiotic-impregnated papers. Plasmids were isolated using the alkaline lysis method as described in Maniatis *et al.*[18] The polymerase chain reaction (PCR) was conducted using a standard PCR kit employing specific forward and reverse primers targeting *mdr* genes, including *bla, tet, acrA, mcr,* and *16s rRNA* genes for 30 cycles (95°C for 45 s, 52°C for 1 min, and 72°C for 1.5 min). The sequencing of PCR fragments was performed at Xcelris Labs Ltd., India. The primers used in this study are presented in Table 1.

2.3. Preparation of phytoextracts and purification of NU2 and CU1

Phytoextracts were prepared by adding 5 mL of ethanol to 1 g of semi-dry chopped root or bark in a 50 mL German-made plastic tube and left overnight at room temperature (25 – 30°C). Purification of NU2 and CU1 phytochemicals was carried out using preparative TLC (20 × 15 cm). The CU1 band was visually identified and cut, while the NU2 band was cut under UV illumination. Silica band was extracted with pure ethanol and centrifuged at 10,000 rpm for 10 min to obtain a clear solution which was then dried at room temperature.

2.4. Biochemical and molecular biological assays

The methyl red assay principle is based on mixed acid fermentation (acetic, lactic, and succinic) by certain bacteria, resulting in a significant pH decrease in the medium, dropping below 4.4. This pH change is indicated by a color change of the pH indicator, methyl red (2-dimethyl-4-amino azobenzene-O-carboxylic acid), turning from yellow when the pH is above 5.1 to red at pH 4.4.

The Voges-Proskauer assay, named after two pioneering microbiologists, is used to detect the formation of acetyl methyl carbinol by bacterial metabolism, a product of the butylene glycol pathway. Organisms such as members of the *Klebsiella-Enterobacter-Hafnia-Serratia* group produce acetoin as the chief end product of glucose metabolism and form smaller quantities of mixed acids. In the presence of atmospheric oxygen and 40% potassium hydroxide, acetoin is converted to diacetyl, which is then converted into a red complex under the catalytic action of alpha-naphthol.

Simmons citrate agar serves as an agar medium used for the differentiation of *Enterobacteriaceae* based on the utilization of citrate as the sole source of carbon. In the early 1920s, Koser developed a liquid medium formulation for the differentiation of fecal coliforms from the coliform group.[19] Simmons later modified this formulation to produce a solid medium that eliminated potential errors when interpreting growth. When the bacteria metabolize citrate, the ammonium salts are broken down to ammonia, which increases alkalinity. This shift in pH turns the bromthymol blue indicator in the medium from green to blue above pH 7.6.

Table 1. Primers used in this study[14]

Name	Sequence of the primers	Tm	Size
P27F	5'-AGA GTT TGA TCC GAA CGC T-3'	62°C	1.4 kb
P1392R	5'-TAC GGC TAC CTT GTT ACG ACT TCA-3'	65°C	
cmrF	5'-TTC GTT AGT CTG CCG TTG CT-3'	56°C	323 bp
cmrR	5'-ATC GCT GGC AAA CAG GGT TA-3'	57°C	
tem-sF1U	5'-ATGATGAGCACYTTTAAAGT-3' Y=C/T	56°C	312 bp
tem-sR1U	5'-TCATTCAGYTCCGKTTCCCA-3'Y=C/T; K=G/T	58°C	
tetF	5'-CTT CGC TAC TTG GAG CCA CT-3'	57°C	910 bp
tetR	5'-GCA GAC AAG GTA TAG GGC GG-3'	57°C	
acrAB-F	5'-ATG CTC TCA GGC AGC TTA GCC-3'	59°C	1 kb
acrAB-R	5'-TGT CAC CAG CCA CTT ATC GCC-3'	59°C	
ctxF1U	5'-AACACMGCMGATAATTCACA-3' M=A/C	59°C	586 bp

The urease assay is simple. Many organisms, especially those that infect the urinary tract, possess a urease enzyme capable of splitting urea in the presence of water to release ammonia and carbon dioxide, thereby increasing the alkalinity of the medium. This alkaline shift causes the indicator phenol red to change from its original orange-yellow color to bright pink.

2.5. High-performance liquid chromatography purification of CU1

The HPLC analysis was conducted at the CSIR-Indian Institute of Chemical Biology, Kolkata, and IIT-Mandi in Himachal Pradesh.[4,5] For the analysis, 5 mg of the TLC-purified active sample was dissolved in 0.5 mL of methanol. After filtration through a membrane filter, 0.1 mL of the sample was loaded onto an HPLC C-18 column pre-equilibrated with methanol.

2.6. Elementary analysis of CU1 and NU2

Elementary analysis was conducted at the Indian Association for the Advancement of Science (IAAS), Kolkata. A total of 4 mg of pure CU1 antibiotic was analyzed for its elemental composition using a Perkin Elmer Elementary Analyzer and compared with the standard. The obtained data included the percentage (%) of carbon and hydrogen. The percentage (%) of oxygen is calculated using Equation I:

$$\%O = 100\% - (\%C + \%H) \tag{I}$$

Our result revealed a notably high oxygen content in CU1 and NU2. In addition, we identified halogen in the structure, confirmed through mass spectroscopy, and further supported by nuclear magnetic resonance (NMR) spectra.

2.7. Mass spectroscopy of CU1

Mass spectroscopy was conducted at the Central Instrument Facility of Bose Institute and the Indian Institute of Science, India. A mass spectrum presents an intensity *versus* m/z (mass-to-charge ratio) plot (histogram), which is unique for each plant alkaloid. Typically, a pure chemical sample is bombarded by a laser, and the resulting positively charged particles are detected by a high-intensity magnet, separating molecular ions and their fragments using a mass spectrometer. This instrument comprises three main components: an ion source, a mass analyzer, and an artificial intelligence (AI)-guided detector. The common fragmentation processes for organic molecules are McLafferty rearrangement and alpha cleavage, which represent unique multiline graphs that aid in identifying a similar molecule and its derivatives. Lighter ions get deflected by the magnetic force more than heavier ions based on Newton's second law of motion, F = ma. In 1911, J. J. Thomson determined the ratio of electrical charge to the mass of an electron (e/m), which is 1811 times less than that of a hydrogen ion.[20] Francis W. Aston introduced the mass spectrograph[21] and won the Nobel Prize in 1922.

2.8. Fourier transform infrared spectroscopy (FTIR) of CU1

The FTIR spectroscopy was conducted at the Central Instrument Facility at Bost Institute, India. The infrared spectra provide information on the functional groups present in a compound. Wave number (v cm^{-1}) is used to measure the infrared absorption within the range of 4000 – 667 cm^{-1} (2.5 – 15 μ wavelength or λ), where v = E/hc and λ=1/v; c = velocity of light and h = Plank's constant. A nonlinear molecule consisting of n atoms has 3n-6 vibrational modes of stretching, rocking, scissoring, wagging, and twisting, offering information on the functional groups of the molecule. Bending vibrations occur at a lower wavenumber than stretching vibrations. Distinctive absorption bands are observed for different types of bonds: carbon triple bond absorption at 2300 – 2000 cm^{-1}; carbon double bond absorption at 1900 – 1500 cm^{-1}; and carbon single bond at 1300 – 800 cm^{-1}; O-H stretching absorption at 3570 cm^{-1}; C-H stretching at 3030 – 2860 cm^{-1}; C-H bending at approximately 1460 cm^{-1}; C=O stretching at approximately 1725 cm^{-1}; N-H stretching at 3500 cm^{-1}; N-H bending at approximately 1650 cm^{-1}; C-N stretching absorption at 1350 cm^{-1}; and C=N at approximately 2200 cm^{-1}. For the analysis, 5 mg HPLC-purified dry active chemical was mixed with 200 mg IR-grade KBr, and a tablet was prepared using a 13-mm die set (Kimaya Engineers, India) at a pressure of 10 kg/cm^2. Spectra were obtained using a Perkin Elmer Spectrum 100 FT-IR Spectrometer (serial no. 80944) for 10 min.[22]

2.9. Nuclear magnetic resonance spectroscopy of CU1

^{3}H-NMR and ^{13}C-NMR analyses were performed at IIT-Mandi, Himachal Pradesh, North India, and Bose Institute. Nuclear magnetic resonance is a spectroscopic technique employed to detect local magnetic moments around odd atomic nuclei when bombarded with radio waves. The most commonly used small molecules are hydrogen (^{1}H) and carbon (^{13}C) but ^{11}B, ^{19}F, ^{23}Na, ^{31}P, ^{35}Cl, etc. also been studied using NMR. At a low energy radio frequency, the nuclei magnetic spin quantum energy is represented by Equation II:

$$E = -\gamma m h B_0 \tag{II}$$

Where B_0 is the field strength, m = Magnetic spin quantum number, γ = Gyromagnetic ratio, and h is Plank's quantum number. For data analysis, NMR absorption spectra were adjusted to the chemical shift (δ) using tetramethylsilane (TMS) as a standard, and the data were expressed as ppm (parts per million). Tetramethylsilane is chemically inert, magnetically isotropic, miscible with most organic solvents, and absorbs at a higher frequency compared to common types of organic protons. Equation III is used to calculate the δ:

$$\delta = \Delta v \times 10^6/\text{oscillator frequency in cps} \quad \text{(III)}$$

Where Δv is the difference in magnetic spin absorption frequencies of the sample and the reference in cps (cps is approximately 700 cycles/s).

2.10. RNA polymerase assay

The RNA polymerase assay was conducted using 0.2 mM XTPs, 10 mM $MgCl_2$, 1 unit RNA of polymerase, 200 ng denatured CT-DNA, and 10 µci α-P^{32}-UTP. The reaction was carried out at 35°C for 15 min, after which it was spotted on diethylaminoethyl (DEAE) paper and washed with 0.5 M sodium phosphate, followed by ethanol. The dried paper was then counted on a scintillation counter. The RNA polymerase assay was performed at the Indian Institute of Science, specifically in Prof. Dipankar Chatterjee's laboratory. A more recent fluorometric RNA polymerase assay using plasmid DNA as a template was conducted at Bose Institute (Dr. Jayanta Mukhopadhyay's laboratory).[15]

3. Results and discussion

3.1. Mechanism of *mdr* genes' function and drug resistance

Figure 1 demonstrates the roles of early penicillinase and beta-lactamases (blaTEM, blaSHV, blaKPC, blaOXA, and blaNDM) in inactivating the penicillin drugs by cleaving the beta-lactam ring to produce penicillinic acid. Similarly, Figure 2 depicts the structure of streptomycin, with arrows indicating different sites on the drugs where AAC and APH-like drug-acetylating and drug-phosphorylating enzymes add a phosphate group or acetyl group to inactivate the drugs. The ANT enzymes also inactivate such drugs by adding an adenyl group to them. In Figure 3, we have presented the classification of different *mdr* genes and their emergence over time. This depiction underscores the rapid evolution of *mdr* genes in response to the introduction of a new antibiotic derivative, conferring upon bacteria the capability to neutralize new drugs. Nowadays, bacterial plasmids acquire DNA nikase, DNA topoisomerases, DNA integrases, transposes, and integrons in such a way that, when exposed to a new chemical, the bacteria begin to rearrange on plasmids to create a new *mdr* gene. Table 2 demonstrates the accumulation of *mdr* genes in both large and small bacterial plasmids. It includes an example of the generation of large plasmids that were sequenced and obtained from the NCBI GenBank Database (www.ncbi.nlm.nih.gov/nucleotide) by putting the accession number (Table 2). We downloaded the plasmid sequences and performed multi-alignment to see the percentage similarities among them because such different plasmids had many common *mdr* genes.[17]

It is challenging to develop a new *mdr* gene against a new antibiotic. However, in the expansive environment of soil, water, and the intestines of humans and animals, nature's laboratories surpass human research laboratories. This is because environmental bacteria are acquiring newer *mdr* genes via conjugation from MDR-bacterial plasmids at a rate of approximately a 5% increase per year. At present, the Ganges River water hosts 45 – 50% ampicillin-resistant bacteria, but the percentage of cefotaxime-resistant bacteria is much lower at 5%. However, imipenem-resistant bacteria are challenging to detect in Ganga River bacteria using conventional plating assays with 0.1 mL water on a 10 cm LB+agar+1 mg/mL imipenem plate. To isolate the imipenem-resistant bacteria, we added 5 µg/mL imipenem to 10 mL Ganges River water and 2 mL 6XLB medium and then incubated the mixture overnight at 37°C. Next, we serially diluted the overnight imipenem-resistant bacteria and plated them on 20 µg/mL imipenem to get single colonies. For simplicity, the total number of bacteria = 12200 cfu/mL, ampicillin-resistant bacteria = 5840 cfu/mL, cefotaxime-resistant bacteria = 50 cfu/mL, and imipenem-resistant

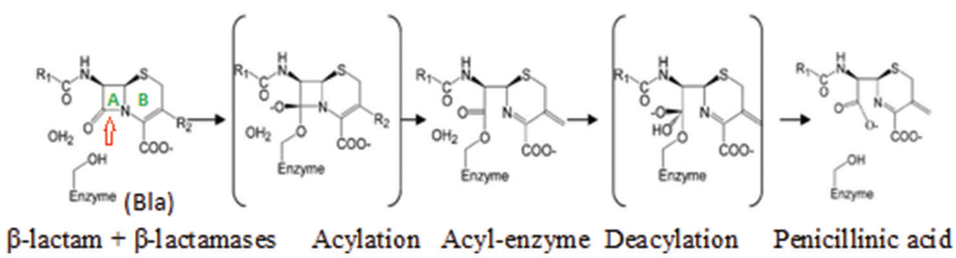


Figure 1. Inactivation of penicillin drugs by beta-lactamases (blaTEM, blaOXA, blaKPC, and blaNDM)

bacteria = 0.2 cfu/mL. Therefore, the chance of infection by imipenem-resistant bacteria is approximately 579 times lower than that of ampicillin-resistant bacteria by bathing or consuming Ganges River water. Thus, more people may be infected by ampicillin-resistant bacteria and still be cured with cefotaxime and imipenem drugs.

Certainly, in 2050, when the percentage of imipenem-resistant bacteria has increased, the situation will be different. It will be simpler to identify imipenem-resistant bacteria by plating 0.1 mL Ganges river water onto a 10 cm LB + agar + imipenem plate. In 2050, doctors are likely to perform drug sensitivity assays first using 100 antibiotic paper disks to determine the nature of a patient's blood, whether it is an MDR, extensively drug-resistant (XDR), or IDR infection. Obtaining such an assay result may cost a few thousand rupees, and one may have to wait for 2 days in order for the doctor to prescribe the correct antibiotic. One of the authors (Asit Kumar Chakraborty) shares a personal experience here. In 2022, the author contrasted an infection, and the doctor prescribed the cefotaxime drug, which was proven effective. However, to avoid the risk of MDR development, the author was subsequently prescribed two higher derivatives of tetracycline and aminoglycoside antibiotics. Fast forward to 2050, and a similar scenario may unfold differently. For instance, if the author were to bathe in the Ganges River, the initial cefotaxime treatment may not be effective, and if amikacin was also ineffective, then hospitalization would be recommended, usually accompanied by drug-sensitivity testing. In such cases, doripenem and meropenem therapy may be considered next, but if the totally drug-resistant (TDR) infection or if all 100 available antibiotics in Kolkata medical stores have failed, then the doctor would seek help from the USA to obtain expensive and high-risk investigational drugs. This is why scientists have predicted that there could be 10 million deaths in Asia annually by 2050. Simply put, people will not be able to afford the costly therapy, leading to their demise. In this paper, we have demonstrated the inactivation mechanism by diverged penicillinases (Figure 1) and have also demonstrated the acetylation, adenylation, and phosphorylation of different aminoglycosides acetyl-transferase (AAC), aminoglycoside adenyl-transferase (AAD), and aminoglycoside phosphotransferase (APH) enzymes (Figure 2). In addition, we have depicted the gradual discovery of new antibiotics with the subsequent generation of new *mdr* genes to inactivate these new antibiotics (Figure 3). This ongoing process has been occurring since 1970 until the present day, and doctors are becoming increasingly exhausted by the continuous development of new MDR bacteria.

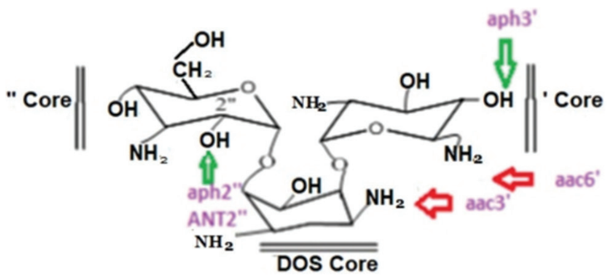


Figure 2. Inactivation of streptomycin by AAC, APH, and ANT MDR enzymes

Abbreviations: AAC: Acetyltransferases; APH: Phosphotransferases; ANT: Adenyltransferases; MDR: Multidrug-resistant.

In Figure 4, we presented the PCR assay results for the blaCTX-M1/2/9 genes using degenerative primers.

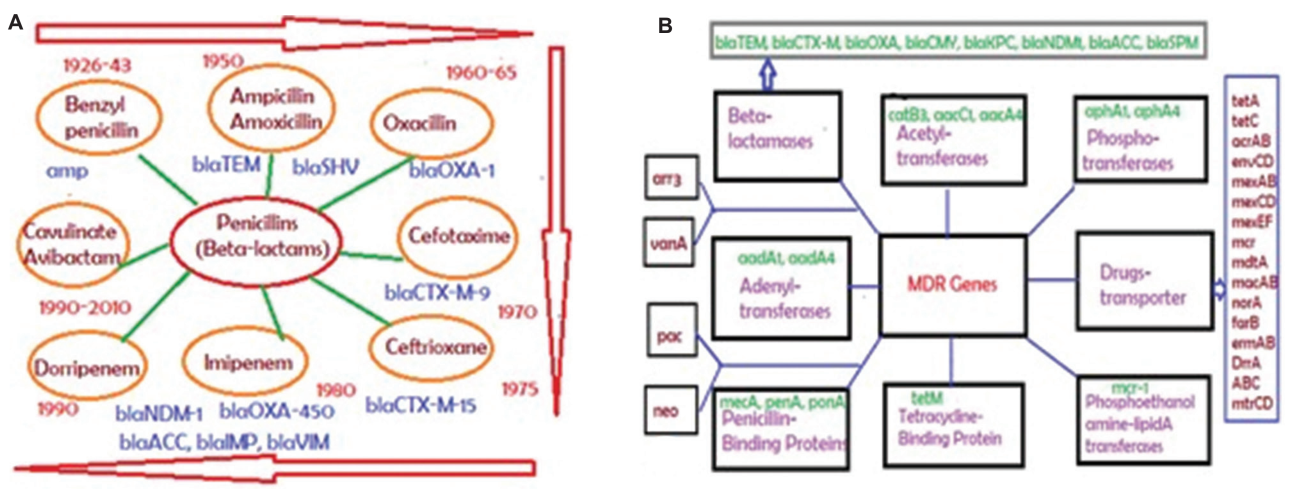


Figure 3. Drugs and MDR genes (A) Gradual discovery of new penicillin drugs with time and the creation of new beta-lactamase derivatives. (B) Name of the different *mdr* genes and so many derivatives for beta-lactamases and the tetracycline membrane transporters, and other drug transporters (acrAB-TolC, mexAB-OprM, and macAB-TolC)

Table 2. Multidrug-resistant (*mdr*) genes in small and large plasmids

Accession number	Size (kb)	Multidrug-resistant marker genes: Drug transporters and antibiotic-inactivating enzymes	GenBank (year)	Pathogenic bacterial name
KC543497	501	blaOXA-10, MFS, blaTEM-8, ble, catB8, and aac3'	2014	*P. aerogenosa*
NC_018107	353	emrE, , crcB, NAT, terA/D, aph*, and blaTEM	2015	*K. oxytoca*
NC_022078	317	ABC, cat, aph*, aac3', cmr, tetA, and blaKPC	2015	*K. pneumoniae*
MG252895	300	FloR, tetA, StrB/A, sul2, blaCMY/OXA/NDM, aac6'-Ib, aphA7, arr2, cmlA1, aphA6, and sul1	2020	*E. coli*
LN555650	299	sul1, strA, catB, blaACC-1, aacA4, and blaVIM-1	2015	*S. enterica*
CP004000	295	bla_SHV-12, blaTEM-1, and ter C/A	2014	*K. pneumoniae*
CP007558	272	blaAmpC, ABC, sul1, blaTEM, aad, and ble	2014	*C. freundii*
JN420336	267	blaNDM1, blaOXA1, aac6', qnrB1, cat, and blaCTX-M,	2020	*K. pneumoniae*
NC_014312	251	blaKPC2, mph 2, ABC, blaAmpC, and qnrB	2014	*K. pneumoniae*
AP012055	250	$blaNDM_1$, aadA2, catA1, and qacA1	2013	*K. pneumoniae*
KM877269	249	aad, floR, hph, aac6'/3', , blaOXA-1, catB, arr3, and sul1	2015	*S. enterica*
CP011634	227	blaOXA, aad, blaTEM, aad, sul1, aac, and blaTEM	2015	*K. oxytoca*
HG530658	223	blaACC-1, strA, aadA2, aac3', rcnA, and pcoS	2015	*E. coli*
NC_019375	180	blaVIM, aacA7, dhfr, ANT3', SHV-5, sul1, and aph 3'	2014	*P. stuartii*
JX442976	172	tetA, aph, sul2, aadA1, blaOXA-10, Qnr, and blaCMY-16	2013	*K. pneumoniae*
NC_022522	168	blaCTX-M25, aacA4*, strB, strA, aadB, and blaOXA-21	2014	*S. enterica*
NC_012692	167	strA, blaCMY2, groEL, stbA, strB, flo^R, and merA	2014	*E. coli*
LN850163	167	MFS, AAA tetA, cat, blaTEM, macB, and blaCTX-M	2015	*E. coli*
NC_019121	166	blaAmpC, sul2, tetA, flo^R, $hygB^R$, and aph 3'	2014	*S. enterica*
LC055503	160	blaSHV12, aac6', blaOXA10, aadA1, sul1, and blaDHA1	2015	*K. pneumoniae*
FJ628167	151	blaKPC, sul1, qnrB4, blaDHA, mph 2, and ABC	2010	*K. pneumoniae*
JX182975	289	Cat, aadA2, sul1, ble, blaNDM1, dhfr, mph 2, and acrA/F	2020	*C. freundii*
KT185451	151	blaTEM/CTXM/SHV12, blaKPC, and blaNDM1	2015	*K. pneumoniae*
KF250428	151	blaIMP-4, aacA4, cmr, and flo^R	2013	*K. pneumoniae*
NC_012690	148	flo^R, tetA, strB, sul2, blaAmpC, sul1, aph, and blaTEM1,	2014	*E. coli*
AP012056	141	Aac3'/6', catB4, tetA, sul2, blaOXA/CTX-M, and strB/A	2013	*K. pneumoniae*
KF954760	140	blaTEM1, strA, strB, and aadA	2014	*K. pneumoniae*
KP893385	137	blaCTXM-65, blaKPC-2, blaSHV-12, and blaTEM-1b	2015	*K. pneumoniae*
HG941719	135	aadA5, mph, blaCTXM/OXA/TEM, aac6', sulI, and tetA	2014	*E. coli*
KF705205	134	hph, strA, aac3'-IV, tetA, and blaTEM-1	2015	*S. enterica*
NC_020087	133	aphA, hph, tetA, $blaLAP_2$, dhfr, ble, and qnrS1	2014	*K. pneumoniae*
CP009115	118	ble, blaOXA-1, qnr, and ble	2014	*K. pneumoniae*
GU256641	110	Sul2, strA, blaTEM1, blaSCO1, aacC2, and blaACC-4	2011	*E. coli*
NC_024978	110	dhfr, aad3, $blaCTX_1$, $EtBr^R$, and ABC-type	2014	*E. coli*
GU256641	110	Sul2, strA, blaTEM, blaSCO-1, aacC2, and blaACC-4	2011	*E. coli*
JX283456	108	blaKPC2, TolA, blaTEM, ABC transporter, and mph 2',	2012	*K. pneumoniae*
JX566770	107	pac, aadA1, dhfrA1, strB, and blaTEM-1	2013	*E. aerogenes*
MF042358	100	Aac3', ble, blaNDM-1, and sul1	2020	*E. cloacae*
CP009116	95	Aph, blaTEM, aac3', MFS, dhfr, aad, arr2, and blaNDM1	2014	*K. pneumoniae*
NC_019889	87	Aac (3')-II, blaNDM-1, sul1, MsrE, and mphE	2014	*K. pneumoniae*
KM406489	87	blaTEM-1	2015	*S. marcescens*

(Cont'd...)

Table 2. ***(Continued)***

Accession number	Size (kb)	Multidrug-resistant marker genes: Drug transporters and antibiotic-inactivating enzymes	GenBank (year)	Pathogenic bacterial name
GU585907	79	aadA2, aphA2, aadA1*, strA, strB, and blaVIM1	2010	*K. pneumoniae*
KF954759	73	blaKPC3, strB, aac (6'), and chrB	2014	*K. pneumoniae*
KJ460501	62	blaCTX-M	2014	*S. sonnei*
NZ_CP008901	52	Dhfr, blaKPC-2	2015	*E. cloacae*
AY046276	51	aadA1, blaOXA-2, sul1, tetA, and ABC	2012	*S. enterica*
KT225462	50	mphE, sul1, blaDHA-1, qbrB, strA, and strB	2015	*K. pneumoniae*
AB61665	47	blaIMP2, aacA4, aadA2, tetA, blaCTX-M, and sul1	2012	*K. pneumoniae*
KJ541071	44	sul1, blaOXA-2, aadA/B, blaTEM, catA1, and blaGES-5	2014	*E. coli*
JX104759	42	blaKPC-2 and ABC	2013	*K. pneumoniae*
KC354802	41	aacA4, aadA1, blaOXA-9, and blaTEM-1	2013	*K. pneumoniae*
NC_021087	26	blaGIM-1, aacA4, aadA1, blaOXA-2, and sul1	2015	*E. cloacae*
JN215524	24	Dhfr, cmlA, blaOXA10, aadA1, qnrB, blaDHA1, and sul1	2012	*C. freundii*
NG_035843	15	blaOXA-30, catB3, arr-3, sul1, qnr, and blaDHA-1	2014	*E. coli*
NG_041456	2.5	blaKPC-2	2014	*P. aeruginosa*
JN677524	1.9	Ble and blaNDM-1	2016	*K. pneumonia*

Abbreviations: *P. aeruginosa: Pseudomonas aeruginosa; K. oxytoca: Klebsiella oxytoca; K. pneumonia: Klebsiella pneumonia; E. coli: Escherichia coli; S. enterica: Salmonella enterica; C. freundii: Citrobacter freundii; P. stuartii: Providencia stuartii; E. aerogenes: Klebsiella aerogenes; E. cloacae: Enterobacter cloacae; S. marcescens: Serratia marcescens; S. sonnei: Shigella sonnei.*

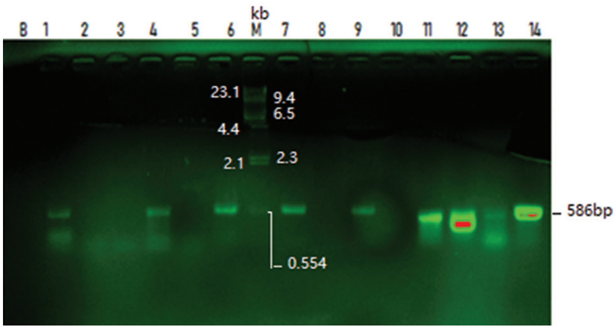


Figure 4. Detection of blaCTX-M1 genes in the Ganges River water MDR bacteria. Among 14 ampicillin- and tetracycline-resistant bacteria, five showed a clear, distinct band of 586 bp. The plasmid DNA was isolated from 6 mL bacteria (four 1.5 mL tubes) by the alkaline-lysis method following RNase A treatment, phenol-chloroform extraction, and ethanol precipitation Notes: M = Lambda Hind-III marker; The primers are CTXF = 5'-AAC, ACM, GCM, GAT, AAT, TCA, CA-3' and CTXR = 5'-CCG, CRA, TAT, CRT, TGG, TGG, TG-3'

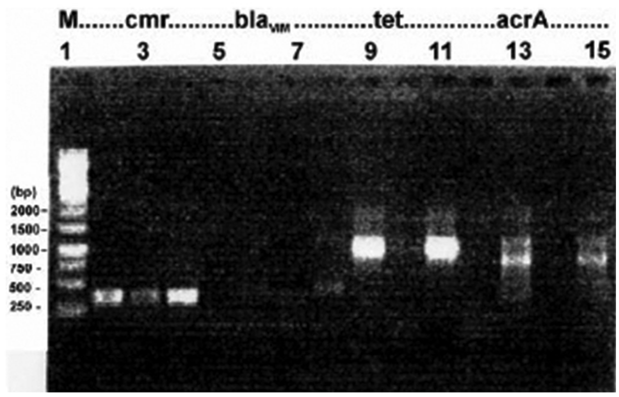


Figure 5. Localization of TET and ACRA drug efflux genes in *Escherichia coli* KT-1_mdr bacteria plasmids. We used both chromosomal DNA and plasmid DNA preparations, and it was prominent that both preparations had *tet* and *acrA* genes. The DNA was not CsCl gradient purified, and thus, due to large plasmids, such chromosome or plasmid classification was not possible. The blaVIM was not there (lanes 5 – 7), and cmr acetyltransferase must be there also (lanes 2 – 4)
Notes: Lane-1=100 bp DNA ladder as molecular weight marker, bla = 519 bp, tet = 910 bp, cmr = 323 bp, and acrA = 1007 bp.

The results revealed that 45% of MDR bacteria isolated from the Ganges River water subjected to ampicillin and tetracycline carried the blaCTX-M gene (586 bp band) in their plasmids. Similarly, the experiment conducted using blaTEM primers demonstrated that 100% of bacteria had the blaTEM gene in their plasmids (data not shown). In Figure 5, we further demonstrated that in MDR bacteria, the acrA and tetC drug efflux genes have been activated, indicating that MDR bacteria can remove multiple antibiotics from their cytoplasm, thereby increasing their MDR.[15]

We have also isolated MDR bacteria from chicken meat, milk, and human hair, which was found to be relatively easy. Plasmids were detected in these various antibiotic-selected MDR bacteria (Figure 6). We isolated 3 kb and 30 kb

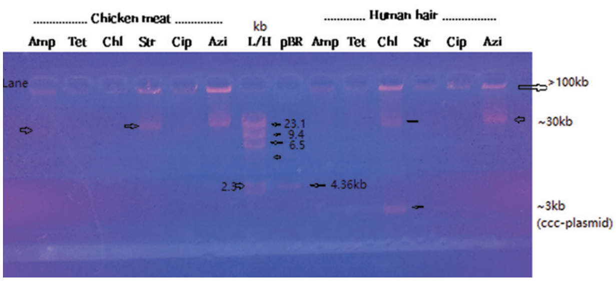


Figure 6. Detection of large plasmids and small plasmids in MDR bacteria selected with six old drugs. Plasmid DNA was isolated from 6 mL bacterial overnight culture and loaded onto 0.8% agarose gel and run for 4 h at 30 volts and ethidium bromide-stained (0.5 μg/mL in 1×TAE buffer). Lambda HindIII DNA and pBR322 plasmid DNA were used as markers. Six bacteria were isolated from a chicken meat sample (Midnapore city), and six other bacteria were isolated from human hair (salon in Kolkata city) and selected with six different antibiotic plates

plasmids and then transfected them into the *Escherichia coli* DH5α laboratory strain that had no *mdr* gene and was highly sensitive to multiple antibiotics, either singly or in combination. We observed ampicillin resistance in the 3 kb plasmid, whereas there was no chloramphenicol resistance in the 30 kb plasmid, suggesting that the *cat* gene was located in larger plasmids (>200 kb), making it challenging to isolate using the conventional method from agarose gel at the Oriental Institute of Science and Technology (OIST) laboratory. Unfortunately, the plasmid contents in *E. coli* KT-1_mdr and *E. coli* KC-1_mdr were high and heterogeneous, resulting in a smear instead of distinct bands. Initially, we attributed this to contamination of our plasmid preparation with chromosomal DNA. Now, we understand that the MDR phenomenon is a mechanism of "genesis within" aimed at protecting bacteria in the intestine from the synthesis of 20 vitamins, which humans cannot produce on their own. Despite advancements, hunger persists, and the world population still does not have access to a balanced diet every day. In addition, many people, especially those from impoverished backgrounds, are reluctant to use multivitamin tablets. Consequently, we continue to depend on intestinal flora or probiotics. All creatures, including humans, depend on nature, and all habitats on Earth live symbiotically in different ways. Thus, it is important to conserve the environment and utilize phytoantibiotics, as proclaimed by the World Health Organization (WHO).

3.2. Isolation of chicken meat, milk, and human hair bacteria and their characterization

Our understanding suggests the presence of MDR bacteria in water, soil, and the intestinal tract. To investigate further, we collected human hair from a salon in South Kolkata, washed it with LB media, and plated it on LB-agar + ampicillin. Similarly, we obtained milk from a local vendor near the Midnapore station area and chicken meat from a local shop in Midnapore city. Antibiotic plates containing ampicillin, cefotaxime, tetracycline, streptomycin, ciprofloxacin, and erythromycin were selected for the analysis of all collected samples.

Next, we presented a few biochemical data supporting the heterogeneity of the MDR bacterial population in chicken meat and human hair. Figure 7 illustrates the results of the methyl red assay; Figure 8 demonstrates the results of Simmon's citrate utilization assay; Figure 9 depicts the urea test results; and Figure 10 demonstrates the sugar utilization test results to indicate the highly heterogeneous population of MDR bacteria. We utilized *E. coli* KT-1_mdr and *P. aeruginosa* DB-2_mdr as standards. The urea test yielded negative results for both MDR bacterial strains, while our selection of bacteria from the six antibiotic plates yielded few positive results. In addition, our 16S rRNA gene sequencing identified *Panalkaligenes* and *Stenotrophonomas* bacteria (Figure 11). However, due to our laboratory's lack of Biosafety Level 3 certification and the identification of potentially pathogenic MDR bacterial isolates in our data, we made the decision to suspend the project.

Recent research by Williams *et al.* involved sequencing the MDR bacterial population isolated from the stools of workers in poultry farms in Bangladesh. They compared the *mdr* genes with those found in poultry cecal and waste water resources.[23] They discovered many *mdr* genes, including *tetQ, blaTEM-1, blaSHV-1, and blaSHV-11*, in human fecal MDR bacteria. The other most abundant *mdr* genes are macrolide-lincosamide-streptogramin-resistant genes. In the poultry cecal samples, however, *SHV-27, blaSHV-110, blaOXA-65, and blaOXA-641*-like mdr genes were also located. In wastewater, the *blaOXA-58 mdr* gene, associated with MDR *Acinetobacter* bacteria, was found to be predominant, and such a protein could hydrolyze carbapenem drugs. The WHO has reported high contamination of poultry meat with CRE Enterobacteriaceae (*E. coli*), *Acinetobacter baumannii,* and *Pseudomonas aeruginosa* with isolated *mdr* genes showing homology to previously described plasmid sequences (accession numbers: CP058135, LC75731, CP054855, KY860573, and MN436715).[24,25]

3.3. Purification of abundant phytochemicals

Our extraction and purification procedures for active chemicals from the ethanol extract using TLC are straightforward, as demonstrated in Figures 12 and 13. We utilized eight 20×15 cm silica gel plates and processed 2.5 – 3 mL extract per TLC step in four tanks with lids. By employing repeated TLC, we were able to concentrate

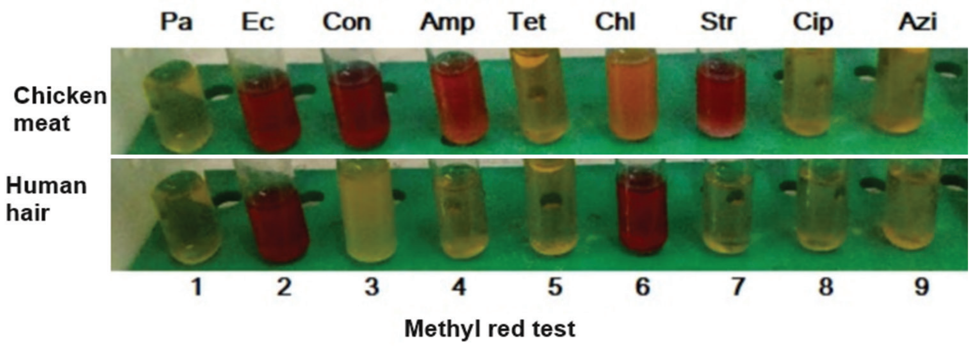


Figure 7. Methyl red assay of different drug-resistant bacteria isolated from chicken meat from Midnapore's meat shop and human hair from a local salon in Kolkata. Ampicillin-selected and streptomycin-selected chicken meat bacteria gave positive tests, whereas only chloramphenicol-selected human hair showed positive results, demonstrating the heterogeneity of the MDR bacterial population

Abbreviations: Pa: *Pseudomonas aeruginosa* BD-2_mdr; Ec: *Escherichia coli* KT-1_mdr standard; Con: Control; Amp: Ampicillin; Tet: Tetracycline; Chl: Chloramphenicol; Str: Streptomycin; Cip: Ciprofloxacin; Azi: Azithromycin.

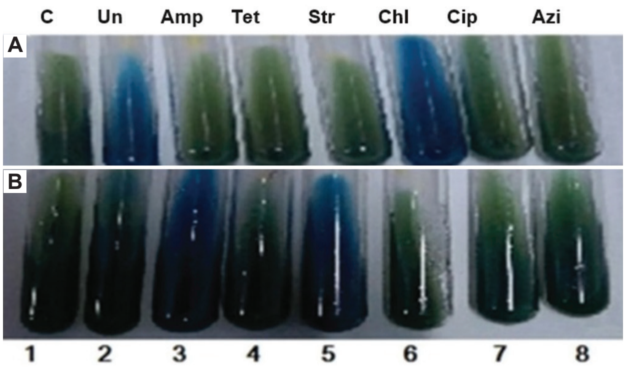


Figure 8. Simmon's citrate utilization test results. Chloramphenicol-selected chicken meat bacteria (A) gave a positive test, while both ampicillin-selected and streptomycin-selected human hair bacteria (B) gave a positive test result (blue color in LB + agar Simmon's citrate slant with bacteria).

Notes: C: Control tube, no bacteria added; Un: Un-selected and total bacteria without antibiotic; amp: Ampicillin; tet: Tetracycline (20 μg/mL); str: Streptomycin (50 μg/mL); chl: Chloramphenicol (25 μg/mL); cip: Ciprofloxacin (50 μg/mL); and azi: Azithromycin (50 μg/mL)

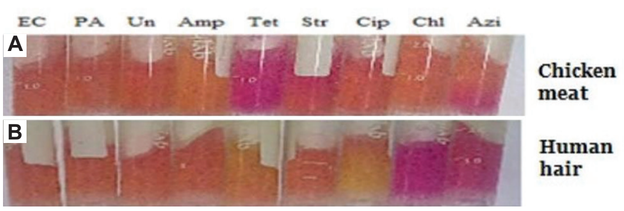


Figure 9. Urease test results. Both *Escherichia coli* and *Pseudomonas aeruginosa* did not yield a positive test. Tetracycline-selected chicken meat bacteria gave a positive result, and chloramphenicol-selected human hair bacteria also gave a positive result

and purify CU1 and NU2 with 90 – 95% purity. Within a week, we collected a sufficient quantity of pure CU1 for biochemical analysis as well as for mass, NMR, and FTIR spectroscopic analyses. Figure 14 presents the potency of

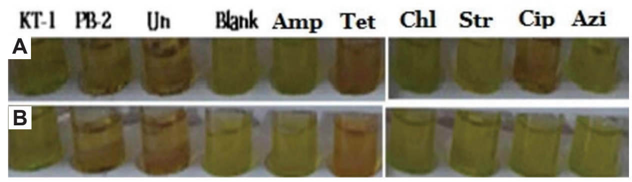


Figure 10. Sugar utilization test results. (A) chicken-derived bacteria and (B) human hair-derived bacteria. KT-1 is multidrug-resistance (MDR) *Escherichia coli,* and PB-2 is MDR *Pseudomonas aeruginosa*. Only tetracycline and ciprofloxacin-selected chicken meat bacteria gave a positive result, while only tetracycline-selected human hair bacteria gave a positive test result. Notes: Un means unselected bacteria; blank means reagent blank

pure CU1, where ampicillin, ciprofloxacin, and ethanol had no effect on the growth of *E. coli* KT-1_mdr bacteria. In Figure 15, we describe the biochemical assays conducted to detect CU1, revealing that it is likely a turpentine polyphenol rather than a glycoside, anthraquinone, or alkaloid.

3.4. Computer-assisted AI-technology-guided modern spectroscopy technology for structure prediction of bio-active phytochemicals

In Figure 16, we presented the mass spectra showing a 75.5 mu line for bromine ion and DBr (82 mu) deviation high molecular weights bands for six bromine atoms. In Figure 17**,** we interpreted the FTIR spectra of CU1, whereby we observed a distinct band at 3000 – 3600 cm^{-1} indicative of the -OH group. In addition, peaks at 2960.1 and 2849.8 cm^{-1} signify $-CH_3$ stretching, while those at 1631.2, 1536.1, 1462.9, or 1387.9 cm^{-1}, suggest O-H bending, likely representing phenolics substituted with bromine atoms at a different position in the benzene ring. Furthermore, peaks at 1259.1 and 1134.0 cm^{-1} indicate C-C-C bending, while a peak at 719.8 cm^{-1} may represent $-CH_2$ rocking. Combining these findings

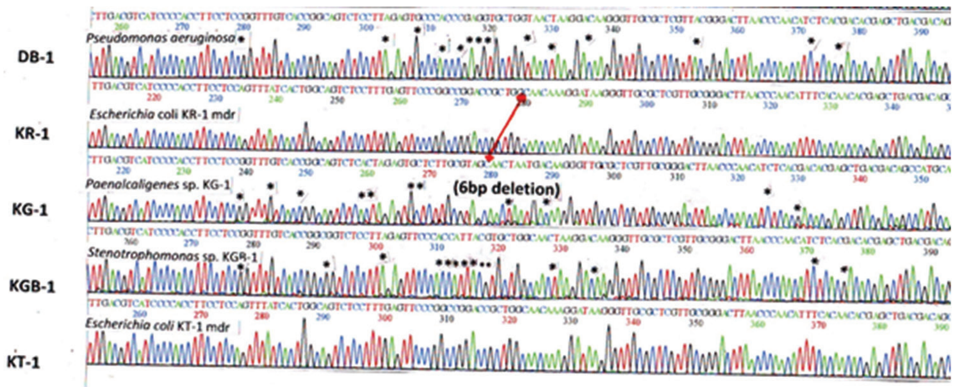


Figure 11. Sanger's DNA sequencing of the 16S rRNA gene of isolated multidrug-resistance bacteria from Ganges River water. Although the rRNA genes of different bacteria had high homology, we found differences among *Escherichia coli, Pseudomonas aeruginosa, Paenalkaligenes* sp., and *Stenotrophomonas* sp. The difference in position with the *E. coli* gene was indicated by black stars. The bacterial name was known from NCBI BLASTN (National Institutes of Health, USA) with derived sequences (not shown here)[14]

Figure 12. Ethanol extraction of *Cassia fistula* bark at room temperature overnight in German-made 50-mL plastic tubes. We avoided grinding the bark mechanically because that caused heating to inactivate the active chemicals CU1 and CU3 of *Cassia fistula* bark

Figure 13. Preparative thin-layer chromatography (20 × 15 cm) of concentrated ethanol extract from *Cassia fistula* bark to isolate CU1 poly-bromo-phenol-saponins (dark band). The CU1 chemical is large but moves fast just below the solvent front (solvent: 40% methanol + 10% acetic acid + 50% water). Ethanol extract 300 – 400 µL was loaded onto each plate, giving a 0.5 cm broad band, dried at room temperature, and put onto 4 tanks with a lid, and ascending chromatography was done for 65 min

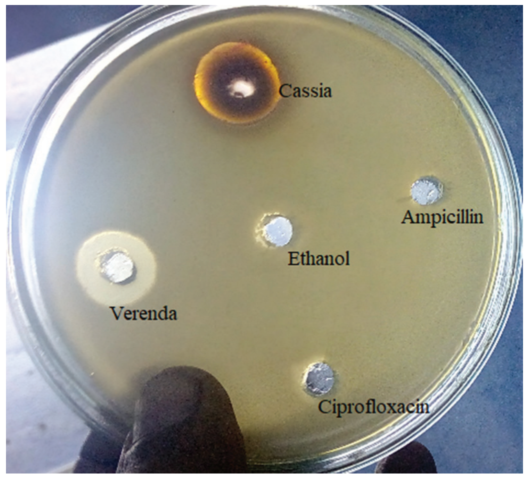


Figure 14. Agar hole assay of different 100% ethanol phytoextracts from *Cassia fistula* bark and *Jatropha gossypifolia* roots

with elementary data analysis revealed that CU1 lacked nitrogen, and its carbon content was determined to be 35.9%, with hydrogen at approximately 5.5%. This data suggested that CU1 is a halogenated derivative, confirmed by the mass spectra. Further, carbon-NMR detected a C-Br bond at 23.7 ppm and a C-O bond, along with a polybenzoid compound at 165 ppm. Proton-NMR further suggested the presence of a polymeric phenol at ό 4.86 – 4.91 ppm with bromine substituents at ό 3.57 – 3.61 ppm (data not shown).[16]

3.5. Animal model and human clinical trial of CU1 MDR-Cure lotion

Next, we assessed the absorption of CU1 in rat intestines and its distribution into the bloodstream to cure *E. coli* KT-1_mdr infection. This treatment cured bacterial systemic infection, and the rats were alive for more

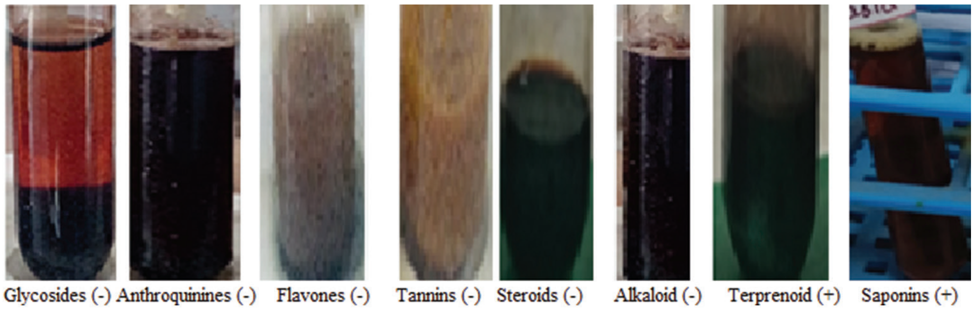


Figure 15. Phytochemicals assays for the thin layer chromatography-purified CU1 from *Cassia fistula* bark ethanol extract

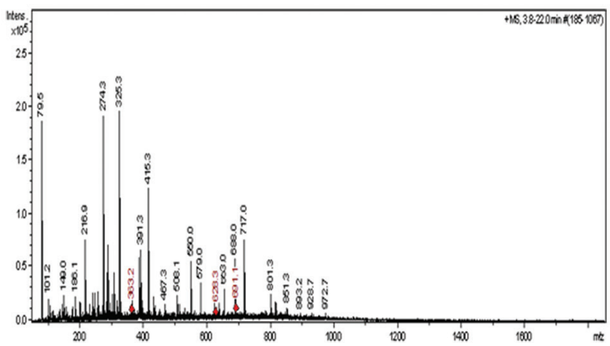


Figure 16. Mass-spectra of CU1 phytochemical. The results reveal a 79.5 mu band corresponding to the Br^- ion. The compound has a molecular weight above 927.7 mu

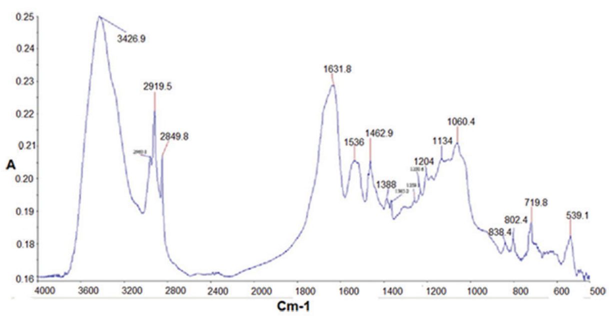


Figure 17. Detection of functional groups of the active compound CU1 by FTIR. The peak at 3426.9 is for N-H stretching and O-H stretching; 2960.1 and 2849.8 cm^{-1} are for CH_3 stretching; 1631.8 and 1536.1 cm^{-1} are for $CO\text{-}NH_2$ scissoring; 1462.9 and 1387.9 cm^{-1} represent O-H bending likely phenolics; 1259.1 and 11.34.0 cm^{-1} for C-C-C bending; and 719.8 may represent $-CH_2$ rocking

Figure 18. Effectiveness of CU1 phytochemicals in rat animal model to clear *Escherichia coli* KT-1_mdr infection. Infection was induced by subcutaneous injection of 0.5 mL bacteria in five different locations on the skin. About 83% of bacterial load in tail-punctured blood was reduced by one oral dose (200 mg) of CU1 and not at all by 0.5 mL cefotaxime (200 mg)

Figure 19. Experiment on human nail chronic multidrug-resistant (MDR) infections. The infection was cured using MDR-Cure phytoextracts

than 3 months (Figure 18). We also treated human nail infection with both CU1 ethanol extract and MDR-Cure phytoextract, comprising neem bark and Haldi rhizome extract (50% ethanolic solution), chosen for their antioxidant and anti-inflammatory properties (Figure 19). Recently, we found that CU1 has no inflammatory effects, and it effectively cures human skin infections (data not shown). However, CU1 has no potential inhibitory effects on the growth of mammalian cells in culture, demonstrating its safety (data not shown). Similarly, we tested CU1 on the growth of molly fishes (red and black), and no significant inhibition of growth was observed, further suggesting the safety of CU1 drug for human use (data not shown).

3.6. Target identification of CU1 as RNA polymerase

We further investigated the molecular target of CU1 and identified its inhibition of *E. coli* RNA polymerase activity. Among the 11 phytochemicals tested, only CU1 exhibited inhibition comparable to the potent bacterial RNA polymerase inhibitor known as rifampicin.[16] As rifampicin is a potent anti-TB drug, we tested the efficacy of the CU1 drug on *Mycobacterium tuberculosis* RNA polymerase, revealing dose-dependent inhibition with increasing concentrations of CU1 (Figure 20).[16] However, in the DNA polymerase assay, there was no inhibition of *E. coli* DNA polymerase by CU1 (Figure 21). CU1 has no potential inhibitory effects on the growth of mammalian cells in culture, revealing that CU1 plays no role in human RNA polymerase activity (data not shown).

3.7. Lantibiotics research has gained momentum with AI-guided and computer-assisted 3D graphics of drug-enzyme interaction

The gramicidin peptide antibiotic has been long recognized to cure skin infections. It is produced by *Bacillus brevis*,

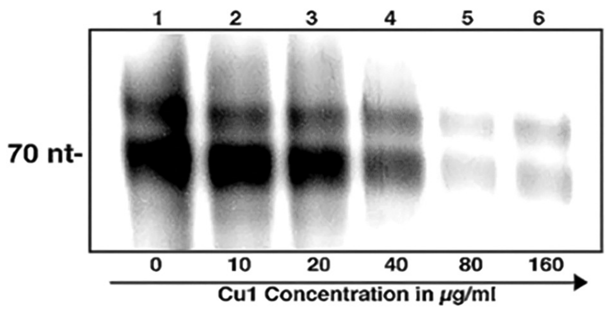


Figure 20. Run on transcription assay using *sinP* plasmid and *Mycobacterium* RNA polymerase. The RNA polymerase activity was gradually inhibited by increasing the concentration of CU1[15]

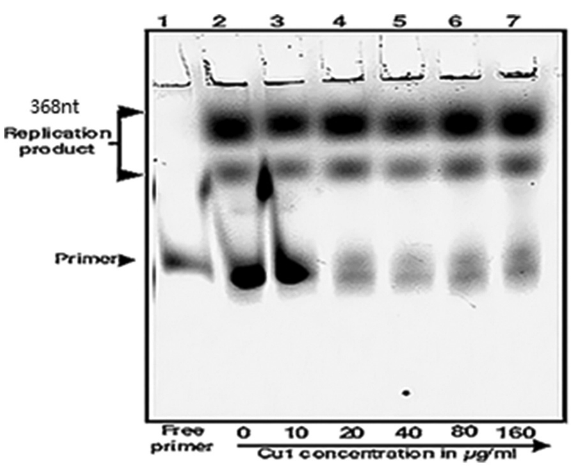


Figure 21. The *in vitro* DNA polymerase assay. The results demonstrate that increasing the concentration of CU1 has no effect on the enzyme

which destroys gram-positive bacteria.[26,27] It disrupts bacterial membrane function and affects DNA and protein structure. Cyclic peptide antibiotics are good and more resistant to inactivation by MDR enzymes. Research is ongoing for the development of salivaricins, nisins, and related lantibiotics for combating XDR tuberculosis (XDR-TB). AI is important in understanding the 3D structural interaction between drugs and target proteins. Computer-assisted modification of drugs generated a good antibiotic with target specificity. For example, 3D crystal structure simulation clearly predicted the rifampicin drug's specificity for the *M. tuberculosis* RNA polymerases and was quite different from other similar RNA polymerases of *E. coli*, *S. aureus,* and *K. pneumoniae*. Thus, rifampicin is a first-line drug despite the emergence of *rpoB* gene (RNA polymerase beta-subunit) mutations that confer resistance. However, other TB-specific drugs such as

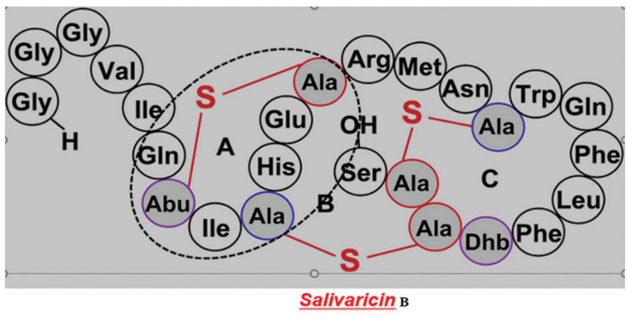


Figure 22. Complex cyclic structures of salivaricin-B lantibiotic that are effective against MDR bacteria. This peptide antibiotic was isolated from the oral bacteria *Streptococcus salivarious*, which contained many large plasmids with 10 salivaricin lantibiotic synthesizing genes

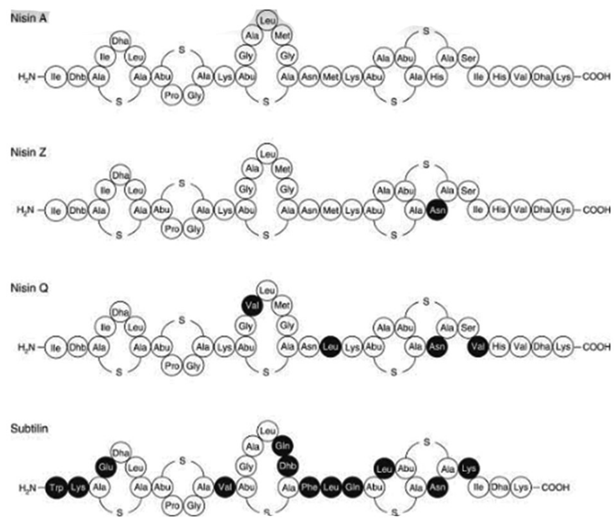


Figure 23. Structures of different isomers of nisin and subtilin lantibiotics. Such structural information was input into a computer device, and with the help of artificial intelligence technology, synthetic lantibiotics with a profound impact on extensively drug-resistant bacteria were formulated

bedaquiline, ethambutol, pyrazinamide, and isoniazid are also prescribed. Therefore, it is crucial to develop more drugs that are resistant to all reported MDR enzymes, such as AMP, CAT, STR, BLA, AAC, APH, AAD, MCR, SUL, and DHFR.[15] The development of more specific TB drugs is essential due to the health risks associated with the current regimen, which entails orally ingesting 12 tablets. AI plays an important role in this endeavor. The mycolic acid complex in the outermost peptidoglycan layer of *M. tuberculosis* is unique to TB bacilli. Their 3D structure formulation is underway using computer graphics to design a more specific type of bedaquiline derivative, enough to kill TB bacilli within 3 – 7 days instead of months.[28,29] AI also aids in predicting easy absorption, blood transport, membrane transport, and target interaction with specificity, whereas 60 thousand cellular proteins would be unaffected. The TB situation in India has deteriorated to the point where it has more cases than any other 250 countries, with 27 million infections in 2022, despite the successful implementation of the directly observed therapy short course (DOTS) program.[7] Modern technologies such as NMR and FTIR, coupled with AI and computer simulation, aid in identifying the correct structure of new drugs, advancing TB treatment efforts.

Using AI technology, many peptide condensation and modification enzymes were cloned into plasmids, overexpressed, and purified. Scientists are developing new cyclic peptide antibiotics *in vitro* by utilizing such enzymes, which are then assayed for their effects on the XDR and TDR clinically isolated bacterial populations.[30] However, this task is challenging and requires increased funding. Unfortunately, these new drugs are likely to be costly, making them inaccessible to people in India as well as in African and Latin American countries due to their financial limitations.[31-33]

In Figures 22 and 23**,** we present the structures of salivaricin B and different nisin lantibiotics. Numerous publications in PubMed highlight the potential use of AI technology and computer software, along with 3D simulation, to develop newer lantibiotics against MDR bacteria.[34-45] In Figure 24**,** we demonstrate the multi-alignment of different lantipeptides that were heterogeneous and strongly diverse, derived from different bacterial origins. However, some similarities were observed in the salivaricins and nisins (Figure 23). In Figure 25**,** we identified a striking similarity between a mutacin and an unidentified protein located in the small plasmid pUA140 of *Streptococcus mutans*, a notorious carcinogen of the oral cavity and teeth. Figures 26 and 27 demonstrate the similarity of peptide condensing and peptide cyclase enzymes between *Streptococcus salivarious* and *Lactobacillus lactis*, both resident oral cavity bacteria. The similarity was almost 50%, and the sequences of those enzymes were well-conserved. Thus, genetic manipulation

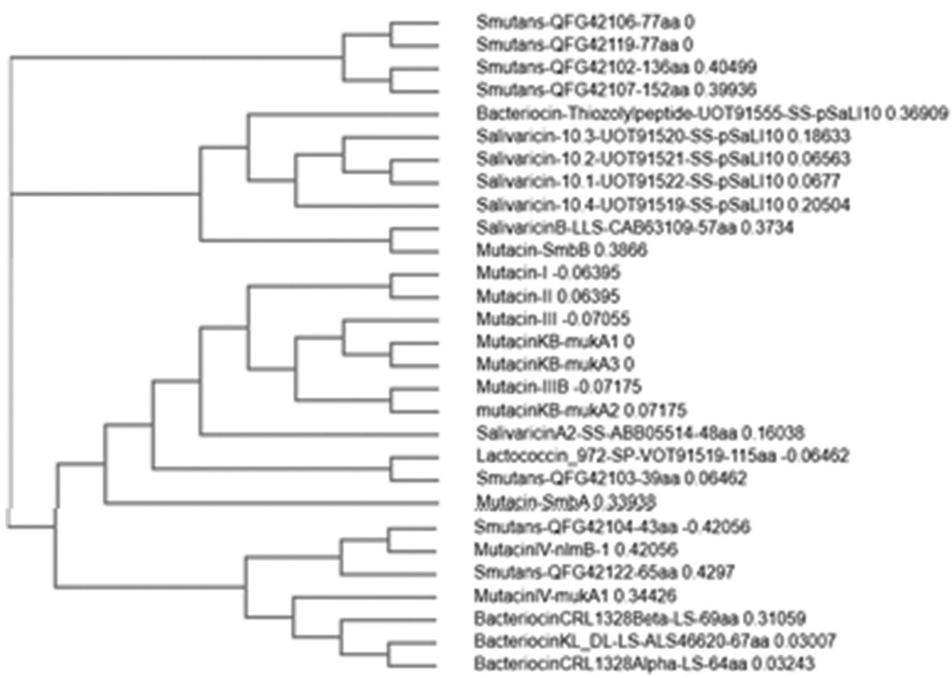


Figure 24. Phylogenetic analysis of different lantibiotics, such as salivaricins, mutacins, and bacteriocins, with broad range activities against multidrug-resistant bacteria

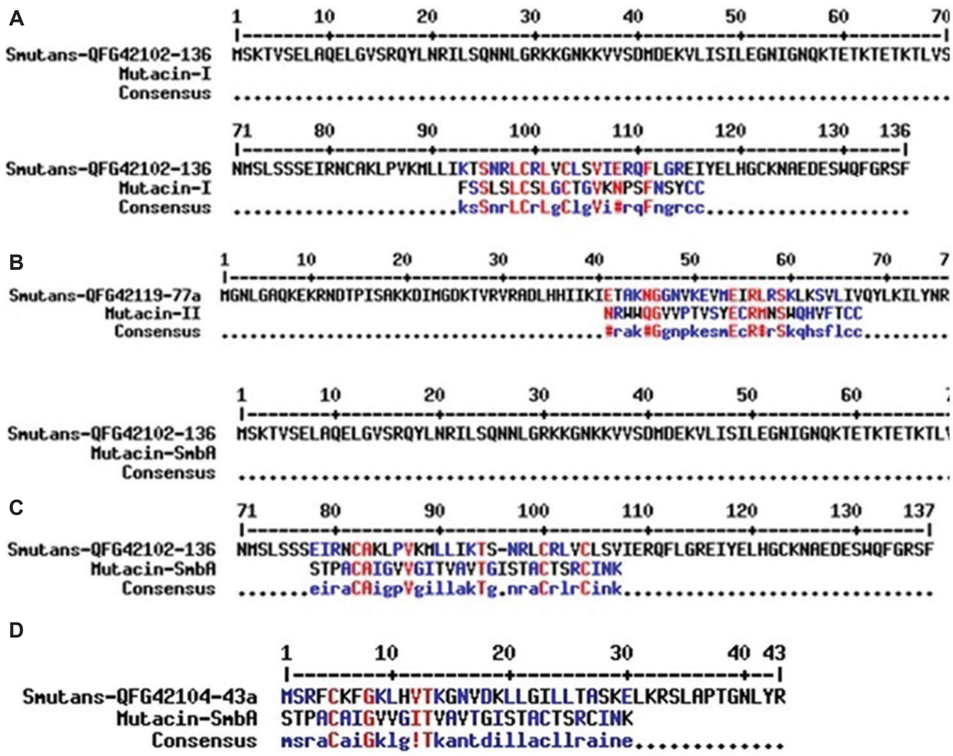


Figure 25. Demonstration of mutacins homology, a pre-lantibiotic protein within the *Streptococcus mutans* plasmid (pUA140). This bacterium, notorious for causing dental caries and oral cancer, utilizes lantibiotic to kill beneficial *Lactobacillus* bacteria. (A) Mutacin-I (strong homology); (B) Mutacin-II; (C) Mutacin-SnbA; (D) another Mutacin-SnbA

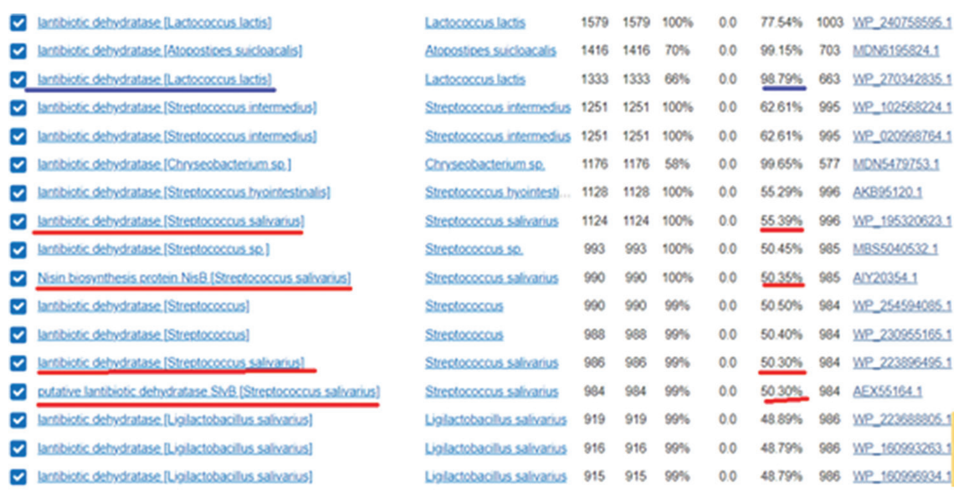
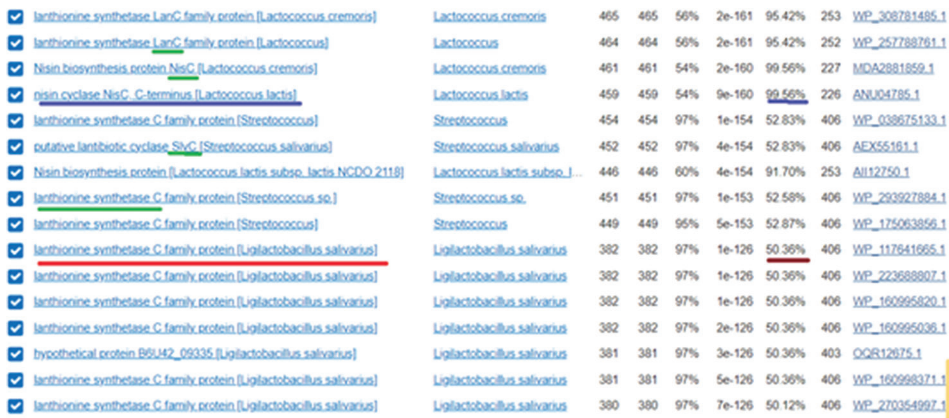


Figure 26. Relationship between lantibiotic genes in oral bacteria, with 50% homology observed between *Streptococcus salivarius* and *Lactobacillus lactis*. These enzymes are cloned, overexpressed, and utilized to develop cyclic peptide antibiotics *in vitro*. The medical AI technology is employed to formulate unique salivaricin and lacticin-like antibiotics targeting membrane structures, inducing leakage of nutrients and subsequent death of extensively drug-resistant bacteria

of these enzymes may lead to novel enzyme activity for the development of novel lantibiotics against XDR-TB.[30]

India has the highest burden of TB, with one in every 550 people at risk of *M. tuberculosis* infection. We trust in herbal drugs and the ancient Indian herbal drug prescription, as documented in Sanskrit books such as the Charaka Samhita, Susruta Samhita, and Atharva Veda, dating back 3000 years.[46,47] However, phytochemical purification



lanthionine synthetase LanC family protein [Lactococcus cremoris]	Lactococcus cremoris	465	465	56%	2e-161	95.42%	253	WP_308781485.1
lanthionine synthetase LanC family protein [Lactococcus]	Lactococcus	464	464	56%	2e-161	95.42%	252	WP_257788761.1
Nisin biosynthesis protein NisC [Lactococcus cremoris]	Lactococcus cremoris	461	461	54%	2e-160	99.56%	227	MDA2881859.1
nisin cyclase NisC, C-terminus [Lactococcus lactis]	Lactococcus lactis	459	459	54%	9e-160	99.56%	226	ANU04785.1
lanthionine synthetase C family protein [Streptococcus]	Streptococcus	454	454	97%	1e-154	52.83%	406	WP_038675133.1
putative lantibiotic cyclase SlvC [Streptococcus salivarius]	Streptococcus salivarius	452	452	97%	4e-154	52.83%	406	AEX55161.1
Nisin biosynthesis protein [Lactococcus lactis subsp. lactis NCDO 2118]	Lactococcus lactis subsp. l...	446	446	60%	4e-154	91.70%	253	AII12750.1
lanthionine synthetase C family protein [Streptococcus sp.]	Streptococcus sp.	451	451	97%	1e-153	52.58%	406	WP_293927884.1
lanthionine synthetase C family protein [Streptococcus]	Streptococcus	449	449	95%	5e-153	52.87%	406	WP_175063856.1
lanthionine synthetase C family protein [Ligilactobacillus salivarius]	Ligilactobacillus salivarius	382	382	97%	1e-126	50.36%	406	WP_117641665.1
lanthionine synthetase C family protein [Ligilactobacillus salivarius]	Ligilactobacillus salivarius	382	382	97%	1e-126	50.36%	406	WP_223688807.1
lanthionine synthetase C family protein [Ligilactobacillus salivarius]	Ligilactobacillus salivarius	382	382	97%	1e-126	50.36%	406	WP_160995820.1
lanthionine synthetase C family protein [Ligilactobacillus salivarius]	Ligilactobacillus salivarius	382	382	97%	2e-126	50.36%	406	WP_160995036.1
hypothetical protein B6U42_09335 [Ligilactobacillus salivarius]	Ligilactobacillus salivarius	381	381	97%	3e-126	50.36%	403	OQR12675.1
lanthionine synthetase C family protein [Ligilactobacillus salivarius]	Ligilactobacillus salivarius	381	381	97%	5e-126	50.36%	406	WP_160998371.1
lanthionine synthetase C family protein [Ligilactobacillus salivarius]	Ligilactobacillus salivarius	380	380	97%	7e-126	50.12%	406	WP_270354997.1

Figure 27. LanC/NisC/SlvC peptide cyclase proteins exhibit 50% homology across oral bacteria. There are many isomers of nisin lantibiotic produced by *Streptococcus lactis*: nisin-A, nisin-Z, nisin-Q, and nisin-U2. Nisin-H was isolated from *Streptococcus hyointestnalis,* and nisin-P from *Streptococcus gallolyticus.* Nisin-O was isolated from a gut bacteria called *Blautia obeum* A2-162. Nisin-J was produced by a Staphylococcus species, and nisin-G was produced by fecal bacteria such as *Streptococcus uberis* and *Streptococcus sius*[52]

methods were not developed then, resulting in suboptimal doses. At present, we are developing phytodrugs (CU1 and NU2) from a few medicinal plants against MDR bacteria, following a few basic points. The phytochemicals must be present in sufficient amounts, typically constituting 30% of the ethanol extract of bark, root, or leaves. Their potency must be demonstrated by a 15 mm-diameter lysis zone or higher in the LB-agar bacterial lysis zone assay using a 1:5 ratio (plant parts: solvent) for overnight extraction at room temperature in a tightly capped plastic tube or bottle. Furthermore, these chemicals must be easily separated by preparative TLC and detectable either by the naked eye or the UV-shadow technique. Such phytochemicals must be cytocidal to at least ten MDR bacteria, initially selected with ten different antibiotics, and be resistant to at least six antibiotics, such as ampicillin, amoxicillin, cefotaxime, tetracycline, amikacin, linezolid, ciprofloxacin, novobiocin, trimoxazole, imipenem, streptomycin, chloramphenicol, erythromycin, azithromycin, lomofloxacin, norfloxacin, and tigecycline.[14,15]

Figure 28 outlines future strategies, emphasizing the potential for automation in drug discovery to revolutionize the development of new antibiotics. In other words, with the help of AI methods, we can outsmart XDR bacteria, which possess a myriad of *mdr* genes and transposons. Together with a slow and expensive antibiotic development pipeline, the proliferation of drug-resistant bacteria drives urgent interest in computational methods that promise to expedite candidate drug discovery.[45,48] Given the urgency of the antimicrobial resistance crisis, we must embrace open science best practices in AI-driven antibiotic discovery to accelerate preclinical research on potent new drugs.[49-53] Ultimately, AI-driven enhancements in drug discovery offer many opportunities

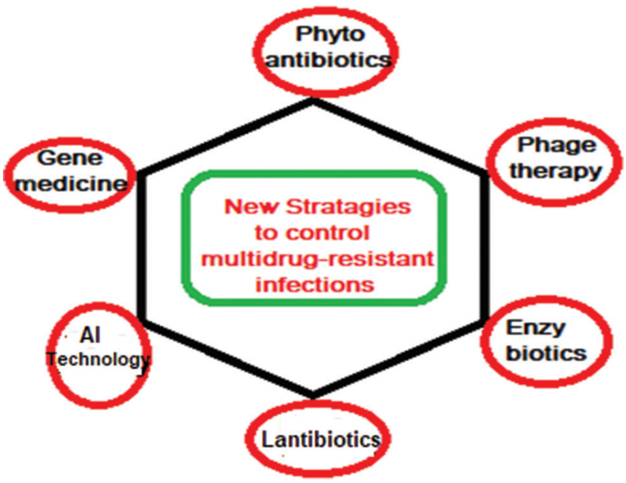


Figure 28. New research dimensions with the help of artificial intelligence technology to control extensively drug resistant and totally drug-resistant bacterial infections as we approach 2050

for future applications in antibiotic development against superbugs.[54-61]

4. Conclusion

The chemical synthesis of antibiotics remains vital, but it has faced challenges in recent times due to high costs and concerns over MDR. Nevertheless, extensive research on novel lantibiotics against XDR bacteria has been reported, positioning peptide antibiotics to take center stage in the coming years. Our research on phytoantibiotics (CU1 and NU2) has gained momentum, with the United Nations recognizing the promising global benefits of such work across all age groups. Our MDR-Cure extract represents a promising antibacterial ayurvedic medicine specifically tailored for skin and nail infections. In addition, while phage therapy has emerged as a potential solution for bacterial infections, our attempts to cultivate a bacteriophage from the drain in Kolkata have been unsuccessful. The use of antibiotics, as well as gene medicine such as antisense and ribozyme, coupled with nanotechnology and AI, maybe a future strategy to combat MDR bacteria. Early intervention is key to averting a projected 10 million deaths by 2050 in Asian, African, and Latin American countries.

Acknowledgments

I am grateful to Dr. Ginger Iv for encouraging me to write this article. This article contains new data on MDR bacteria characterization and novel findings on the NU2 drug, which have been set aside due to patent application. As a retired professor of Biochemistry from Vidyasagar University, I had the privilege of collaborating with Meghna Maity and Sumana Sahoo on their MSc Biochemistry dissertations, which focused on characterizing MDR bacteria from chicken meat and human hair. I am also thankful to Dr. Jayanta Mukhopadhyay of Bost Institute, Kolkata, for his help in the article writing.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Asit Kumar Chakraborty
Formal analysis: Asit Kumar Chakraborty
Investigation: Meghna Maity, Sumana Sahoo
Methodology: Meghna Maity, Sumana Sahoo
Writing – original draft: Asit Kumar Chakraborty
Writing – review & editing: Asit Kumar Chakraborty

Ethics approval and consent to participate

The animal experiment was approved by the Institutional Ethics Committee of Vidyasagar University .

Consent for publication

Asit Kumar Chakraborty gave consent to publish his data in this study.

Availability of data

Data used in this work is available from the corresponding author upon reasonable request.

Further disclosure

Asit Kumar Chakraborty gave online lectures on phytodrugs against MDR-bacteria aspect in Pharmaceutical Congress, Aver Conference, Tokyo on March 20, 2023; Clinical Microbiology conference, London dated March 21, 2023; Biotechnology conference, Rome on June 14, 2023, Pharmaceutical conference, Valencia, Spain on September 14, 2023, and Natural Medicine congress, Chicago, USA on October 11, 2023. A related review and research article by Asit Kumar Chakraborty relating MDR-TB was deposited to BioRxiv preprint (Doi: https://doi.org/10.1101/2023.09.04.556143 and https://www.biorxiv.org/content/10.1101/2020.11.04.369058v1). Most of the data were published in new open-access journals between 2015 and 2023.

References

1. McArthur AG, Waglechner N, Nizam F, *et al*. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother*. 2013;57(7):3348-3357.

   doi: 10.1128/AAC.00419-13

2. Woappi Y, Gabani P, Singh A, Singh OV. Antibiotrophs: The complexity of antibiotic-subsisting and antibiotic-resistant microorganisms. *Crit Rev Microbiol*. 2016;42(1):17-30.

   doi: 10.3109/1040841X.2013.875982

3. Das B, Verma J, Kumar P, Ghosh A, Ramamurthy T. Antibiotic resistance in *Vibrio cholerae*: Understanding the ecology of resistance genes and mechanisms. *Vaccine*. 2020;38(Suppl 1):A83-A92.

   doi: 10.1016/j.vaccine.2019.06.031

4. Chakraborty AK. High mode contamination of multi-drug resistant bacteria in Kolkata: Mechanism of gene activation and remedy by heterogeneous phyto-antibiotics. *Indian J Biotechnol*. 2015;14:149-159.

5. Chakraborty AK. Ganga action plan, heterogeneous phyto-antibiotics and phage therapy are the best hope for India tackling superbug spread and control. *Indian J Biol Sci*. 2017;23:34-51.

6. Moo CL, Yang SK, Yusoff K, *et al*. Mechanisms of Antimicrobial Resistance (AMR) and alternative approaches to overcome AMR. *Curr Drug Discov Technol*. 2020;17(4):430-447.

   doi: 10.2174/1570163816666190304122219

7. Chakraborty AK. Current status and unusual mechanism of multiresistance in *Mycobacterium tuberculosis*. *J Health Med Inform*. 2019;10(1):328.

   doi: 10.4172/2157-7420.1000328

8. Barbour A, Tagg J, Abou-Zied O, Philip K. New insights into the mode of action of the lantibiotic salivaricin B. *Sci Rep*. 2016;6:31749.

   doi: 10.1038/srep31749

9. Barbour A, Wescombe PA, Simtj L. Evolution of lantibiotic salivaricins: New weapons to fight infectious diseases.

Trends Microbiol. 2020;28(7):578-593.

doi: 10.1016/j.tim.2020.03.001

10. Miklasińska-Majdanik M. Mechanisms of resistance to macrolide antibiotics among *Staphylococcus aureus*. *Antibiotics* (*Basel*). 2021;10(11):1406.

doi: 10.3390/antibiotics10111406

11. Cowan MM. Plant products as antimicrobial agents. *Clin Microbiol Rev*. 1999;12:564-582.

doi: 10.1128/CMR.12.4.564

12. Ren Y, Yu J, Kinghorn AD. Development of anticancer agents from plant-derived sesquiterpene lactones. *Curr Med Chem*. 2016;23(23):2397-2420.

doi: 10.2174/0929867323666160510123255

13. Daglia M. Polyphenols as antimicrobial agents. *Curr Opin Biotechnol*. 2011;23:174-181.

doi: 10.1016/j.copbio.2011.08.007

14. Chakraborty AK. Multi-drug resistant bacteria from Kolkata Ganga River with heterogeneous MDR genes have four hallmarks of cancer cells but could be controlled by organic phyto-extracts. *Biochem Biotechnol Res*. 2017;5(1):11-23.

15. Chakraborty AK, Saha S, Poria K, Samanta T, Gautam S, Mukhopadhyay J. A saponin-polybromophenol antibiotic (CU_1) from *Cassia fistula* bark against multi-drug resistant bacteria targeting RNA polymerase. *Curr Res Pharmacol Drug Discov*. 2022;3:100090.

doi: 10.1016/j.crphar.2022.100090

16. Liu K, Huigens RW 3rd. Instructive advances in chemical microbiology inspired by nature's diverse inventory of molecules. *ACS Infect Dis*. 2020;6(4):541-562.

doi: 10.1021/acsinfecdis.9b00413

17. Tan HM, Lall AC, Keppo J, Chen SL. Evaluation of a new antiresistic strategy to manage antibiotic resistance. *J Glob Antimicrob Resist*. 2023;33:368-375.

doi: 10.1016/j.jgar.2023.03.006

18. Maniatis T, Fritsch EF, Sambrook J. *Molecular Cloning-A Laboratory Manual.* Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press; 1982.

19. Koser SA. Correlation of Citrate utilization by members of the colon-aerogenes group with other differential characteristics and with habitat. *J Bacterol.* 1924;9:59-77.

doi: 10.1128/jb.9.1.59-77.1924

20. Thomson JJ. Cathode rays. *Philos Mag.* 1897;44(269):293-316.

doi: 10.1080/14786449708621070

21. Aston FW. LXXIV. A positive ray spectrograph. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. Vol. 38. United States: Creative Media Partners, LLC; 1919. p. 707-714.

doi: 10.1080/14786441208636004

22. Song Y, Cong Y, Wang B, Zhang N. Applications of Fourier transform infrared spectroscopy to pharmaceutical preparations. *Expert Opin Drug Deliv.* 2020;17(4):551-571.

doi: 10.1080/17425247.2020.1737671

23. Willams AD, Rousham E, Neal AL, *et al*. Impact of contrasting poultry exposures on human, poultry, and wastewater antibiotic resistomes in Bangladesh. *Microbiol Spectr*. 2023;11:e01763-23.

doi: 10.1128/spectrum.01763-23

24. Carvalho I, Chenouf NS, Carvalho JA, *et al*. Multidrug-resistant *Klebsiella pneumoniae* harboring extended spectrum β-lactamase encoding genes isolated from human septicemias. *PLoS One*. 2021;16(5):e0250525.

doi: 10.1371/journal.pone.0250525

25. Guclu AU, Gozen AG. Genetic diversity of OXA-like genes in multidrug-resistant *Acinetobacter baumannii* strains from ICUs. *Clin Lab*. 2020;66(10):20215-2019.

doi: 10.7754/Clin.Lab.2020.200135

26. Hotchkiss RD, Dubos RJ. Fractionation of the bactericidal agent from cultures of a soil *Bacillus*. *J Biol Chem*. 1940;132:791-792.

27. Ganz T. Defensins: Antimicrobial peptides of innate immunity. *Nat Rev Immunol*. 2003;3:710-720.

doi: 10.1038/nri1180

28. Imran M, Abida, Alotaibi NM, *et al*. Computer-assisted discovery of safe and effective DprE1/aaRSs inhibitors against TB utilizing drug repurposing approach. *J Infect Public Health*. 2023;16(4):554-572.

doi: 10.1016/j.jiph.2023.02.005

29. Taira J, Nagano T, Kitamura M, Yamaguchi M, Sakamoto H, Aoki S. Structural modification of a novel inhibitor for mycobacterium enoyl-acyl carrier protein reductase assisted by *in silico* structure-based drug screening. *Int J Mycobacteriol.* 2020;9(1):12-17.

doi: 10.4103/ijmy.ijmy_184_19

30. Wayah SB, Philip K. Purification, characterization, mode of action, and enhanced production of salivaricin MMAYE1, a novel bacteriocin from *Lactobacillus salivarius* SPW1 of human gut origin. *Electron J Biotechnol.* 2018;35:39-47.

doi: 10.1016/j.ejbt.2018.08.003

31. Foreman KJ, Marquez N, Dolgert A, *et al*. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: Reference and alternative scenarios for 2016-40 for 195 countries and territories. *Lancet*. 2018;392(10159):2052-2090.

doi: 10.1016/S0140-6736(18)31694-5

32. Allel K, Day L, Hamilton A, *et al*. Global antimicrobial-

resistance drivers: An ecological country-level study at the human-animal interface. *Lancet Planet Health*. 2023;7(4):e291-e303.

doi: 10.1016/S2542-5196(23)00026-8

33. Tanimura T, Jaramillo E, Weil D, Raviglione M, Lönnroth K. Financial burden for tuberculosis patients in low-and middle-income countries: A systematic review. *Eur Respir J*. 2014;43(6):1763-1775.

    doi: 10.1183/09031936.00193413

34. Phan J, Nair D, Jain S, *et al*. Alterations in gut microbiome composition and function in irritable bowel syndrome and increased probiotic abundance with daily supplementation. *mSystems*. 2021;6(6):e01215-21.

    doi: 10.1128/mSystems.01215-21

35. De Gaetano GV, Lentini G, Famà A, Coppolino F, Beninati C. Antimicrobial resistance: Two-component regulatory systems and multidrug efflux pumps. *Antibiotics* (*Basel*). 2023;12(6):965.

    doi: 10.3390/antibiotics12060965

36. Alock BP, Raphenya AR, Lau TTY, *et al*. CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res*. 2020;48:D517-D525.

    doi: 10.1093/nar/gkz935

37. Barbour A, Smith L, Oveisi M, *et al*. Discovery of phosphorylated lantibiotics with proimmune activity that regulate the oral microbiome. *Proc Natl Acad Sci U S A*. 2023;120(22):e2219392120.

    doi: 10.1073/pnas.2219392120

38. Pei ZF, Zhu L, Sarksian R, van der Donk WA, Nair SK. Class V Lanthipeptide cyclase directs the biosynthesis of a stapled peptide natural product. *J Am Chem Soc*. 2022;144(38):17549-17557.

    doi: 10.1021/jacs.2c06808

39. Tovillas P, Navo CD, Oroz P, *et al*. Synthesis of β2,2-amino acids by stereoselective alkylation of isoserine derivatives followed by nucleophilic ring opening of quaternary sulfamidates. *J Org Chem*. 2022;87(13):8730-8743.

    doi: 10.1021/acs.joc.2c01034

40. Bothwell IR, Caetano T, Sarksian R, Mendo S, van der Donk WA. Structural analysis of class I Lanthipeptides from *Pedobacter lusitanus* NL19 reveals an unusual ring pattern. *ACS Chem Biol*. 2021;16(6):1019-1029.

    doi: 10.1021/acschembio.1c00106

41. Joaquin D, Lee MA, Kastner DW, *et al*. Impact of dehydroamino acids on the structure and stability of incipient 3_{10}-helical peptides. *J Org Chem*. 2020;85(3):1601-1613.

    doi: 10.1021/acs.joc.9b02747

42. De Luca S, Digilio G, Verdoliva V, Tovillas P, Jiménez-Osés G, Peregrina JM. Lanthionine peptides by S-alkylation with substituted cyclic sulfamidates promoted by activated molecular sieves: Effects of the sulfamidate structure on the yield. *J Org Chem*. 2019;84(22):14957-14964.

    doi: 10.1021/acs.joc.9b02306

43. Dickman R, Mitchell SA, Figueiredo AM, Hansen DF, Tabor AB. Molecular recognition of lipid II by Lantibiotics: Synthesis and conformational studies of analogues of nisin and mutacin rings A and B. *J Org Chem*. 2019;84(18):11493-11512.

    doi: 10.1021/acs.joc.9b01253

44. Chen H, Zhang Y, Li QQ, Zhao YF, Chen YX, Li YM. De novo design to synthesize lanthipeptides involving cascade cysteine reactions: SapB Synthesis as an example. *J Org Chem*. 2018;83(14):7528-7533.

    doi: 10.1021/acs.joc.8b00259

45. Ma C, Peng Y, Li H, Chen W. Organ-on-a-Chip: A new paradigm for drug development. *Trends Pharmacol Sci*. 2021;42(2):119-133.

    doi: 10.1016/j.tips.2020.11.009

46. Najmi A, Javed SA, Al Bratty M, Alhazmi HA. Modern approaches in the discovery and development of plant-based natural products and their analogues as potential therapeutic agents. *Molecules*. 2022;27(2):349.

    doi: 10.3390/molecules27020349

47. Sivadas N, Kaul G, Akhir A, *et al*. Naturally derived malabaricone B as a promising bactericidal candidate targeting multidrug-resistant *Staphylococcus aureus* also possess synergistic interactions with clinical antibiotics. *Antibiotics* (*Basel*). 2023;12(10):1483.

    doi: 10.3390/antibiotics12101483

48. Torres MT, de la Fuente-Nunez C. Toward computer-made artificial antibiotics. *Curr Opin Microbiol*. 2019;51:30-38.

    doi: 10.1016/j.mib.2019.03.004

49. Torres MDT, Cao J, Franco OL, Lu TK, de la Fuente-Nunez C. Synthetic biology and computer-based frameworks for antimicrobial peptide discovery. *ACS Nano*. 2021;15(2):2143-2164.

    doi: 10.1021/acsnano.0c09509

50. Aronica PGA, Reid LM, Desai N, *et al*. Computational methods and tools in antimicrobial peptide research. *J Chem Inf Model*. 2021;61(7):3172-3196.

    doi: 10.1021/acs.jcim.1c00175

51. Gray DA, Wenzel M. Multitarget approaches against multiresistant superbugs. *ACS Infect Dis*. 2020;6(6):1346-1365.

    doi: 10.1021/acsinfecdis.0c00001

52. Fields FR, Lee SW, McConnell MJ. Using bacterial genomes and essential genes for the development of new antibiotics. *Biochem Pharmacol*. 2017;134:74-86.

doi: 10.1016/j.bcp.2016.12.002

53. Rana R, Awasthi R, Sharma B, Kulkarni GT. Nanoantibiotic formulations to combat antibiotic resistance - old wine in a new bottle. *Recent Pat Drug Deliv Formul*. 2019;13(3):174-183.

doi: 10.2174/1872211313666190911124626

54. Munir MU, Ahmed A, Usman M, Salman S. Recent advances in nanotechnology-aided materials in combating microbial resistance and functioning as antibiotics substitutes. *Int J Nanomedicine*. 2020;15:7329-7358.

doi: 10.2147/IJN.S265934

55. Manrique PD, López CA, Gnanakaran S, Rybenkov VV, Zgurskaya HI. New understanding of multidrug efflux and permeation in antibiotic resistance, persistence, and heteroresistance. *Ann N Y Acad Sci*. 2023;1519(1):46-62.

doi: 10.1111/nyas.14921

56. Mulat M, Pandita A, Khan F. Medicinal plant compounds for combating the multi-drug resistant pathogenic bacteria: A review. *Curr Pharm Biotechnol*. 2019;20(3):183-196.

doi: 10.2174/1872210513666190308133429

57. Tarín-Pelló A, Suay-García B, Pérez-Gracia MT. Antibiotic resistant bacteria: Current situation and treatment options to accelerate the development of a new antimicrobial arsenal. *Expert Rev Anti Infect Ther*. 2022;20(8):1095-1108.

doi: 10.1080/14787210.2022.2078308

58. Lv S, Wang Y, Jiang K, *et al*. Genetic engineering and biosynthesis technology: Keys to unlocking the chains of phage therapy. *Viruses*. 2023;15(8):1736.

doi: 10.3390/v15081736

59. Zeituni EM, Raterman EL. NIAID's comprehensive support mechanisms for antibiotic development. *ACS Infect Dis*. 2020;6(6):1299-1301.

doi: 10.1021/acsinfecdis.0c00099

60. Lluka T, Stokes JM. Antibiotic discovery in the artificial intelligence era. *Ann N Y Acad Sci*. 2023;1519(1):74-93.

doi: 10.1111/nyas.14930

61. Xavier BB, Das AJ, Cochrane G, *et al*. Consolidating and exploring antibiotic resistance gene data resources. *J Clin Microbiol*. 2016;54(4):851-859.

doi: 10.1128/JCM.02717-15

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Efficient schema-less text-to-SQL conversion using large language models

Youssef Mellah*, Veysel Kocaman, Hasham Ul Haq, and David Talby

John Snow Labs, Coastal Highway, Lewes, Delaware, United States of America

Abstract

Large language models (LLMs) are increasingly being applied to several tasks including text-to-SQL (the process of converting natural language to SQL queries). While most studies revolve around training LLMs on large SQL corpora for better generalization and then perform prompt engineering during inference, we investigate the notion of training LLMs for schema-less prompting. In particular, our approach uses simple natural language questions as input without any additional knowledge about the database schema. By doing so, we demonstrate that smaller models paired with simpler prompts result in considerable performance improvement while generating SQL queries. Our model, based on the Flan-T5 architecture, achieves logical form accuracy (LFA) of 0.85 on the MIMICSQL dataset, significantly outperforming current state-of-the-art models such as Defog-SQL-Coder, GPT-3.5-Turbo, LLaMA-2-7B and GPT-4. This approach reduces the model size, lessening the amount of data and infrastructure cost required for training and serving, and improves the performance to enable the generation of much complex SQL queries.

Keywords: Large language models; MIMICSQL; Schema-less; Logical form accuracy; Defog-SQL-Coder; GPT-3.5-Turbo; LLaMA-2-7B; GPT-4

***Corresponding author:**
Youssef Mellah
(youssef@johnsnowlabs.com)

Citation: Mellah Y, Kocaman V, Haq HU, Talby D. Efficient schema-less text-to-SQL conversion using large language models. *Artif Intell Health.* 2024;1(2): 96-106. doi: 10.36922/aih.2661

Received: January 6, 2024

Accepted: February 23, 2024

Published Online: April 4, 2024

1. Introduction

Text-to-SQL technology has gained considerable attention in recent years, emerging as a transformative tool for database interaction. Its key advantage lies in enabling users, particularly those with limited SQL knowledge, to use a fine-tuned large language model (LLM) to interact with databases using natural language. This innovation significantly reduces the necessity to learn SQL for data retrieval and analytics from tabular datasets. The effectiveness of such systems hinges on two main aspects: the intuitiveness of usage and the accuracy of the generated queries. Essentially, this means that user prompts should be straightforward and the corresponding SQL queries must accurately address the user's query with high precision.

The growing abundance of structured and semi-structured data in various domains, ranging from e-commerce to healthcare, highlights the importance of the text-to-SQL task. This task gains relevance as the demand for more intuitive interfaces to query and extract information from these databases increases. Traditional SQL queries, which require understanding of both database schema and query syntax, are often challenging for users lacking technical expertise. Text-to-SQL aims to mitigate this challenge by enabling users to formulate queries in natural language, thereby lowering the barriers to data access and analysis.

In the past decade, the field of natural language processing (NLP), especially through the development of LLMs, has seen remarkable progress, substantially enhancing text-to-SQL systems' performance.[1,2] Models such as T5, LLaMA, GPT-3, GPT-3.5, and GPT-4 have been pivotal in advancing natural language understanding and generation, displaying a profound ability to process and produce human-like text. Despite these advancements, adapting these versatile models for specific applications, such as generating SQL queries for structured data, remains a significant challenge.

In this research, we aim to tackle the dual challenges of simplifying input prompts and elevating accuracy in the generation of SQL queries, with a specific focus on the intricate landscape of the medical domain. Given the critical importance of precision in data retrieval within healthcare contexts, our primary goal is to fine-tune Flan-T5-based models using text-to-SQL query pairs meticulously tailored for the medical MIMICSQL dataset.[3] The decision to utilize a medical dataset in our research is driven by the distinctive challenges and precision requirements inherent in health-care data retrieval. The choice of the MIMICSQL dataset, derived from the widely-used MIMIC-III database, provides a realistic and clinically relevant context, allowing us to address the complexities of real-world medical scenarios. Focusing on the medical domain enables us to tailor our approach to the unique intricacies of healthcare data, contributing directly to advancements in medical data management. By enhancing the accuracy of SQL query generation in this specific context, our research seeks to deliver a meaningful impact on the efficiency of data retrieval in medical databases, benefiting health-care professionals, researchers and decision-makers.

To guide our investigation effectively, we pose the following research questions:

(i) How can we optimize the formulation of input prompts to simplify the querying process while maintaining the necessary specificity required for medical data retrieval?
(ii) What adjustments and enhancements can be made to Flan-T5-based models to improve their accuracy in generating SQL queries tailored to the nuances of the medical MIMICSQL dataset?
(iii) How do schema-less questions contribute to streamlining input prompt complexity, and what impact does this simplification have on the overall performance of SQL query generation in the medical domain?

As we delve into these research questions, our methodology strategically leverages schema-less questions, with a deliberate focus on mitigating the challenges posed by complex and lengthy input prompts. While acknowledging that this approach may potentially limit generalization across diverse database schemas, we anticipate that the pronounced enhancement in overall performance will substantiate this deliberate trade-off.

The organization of this paper is as follows: Section 2 presents a thorough review of the existing literature in the text-to-SQL field. Section 3 describes our methodology, including details about the MIMICSQL dataset, preprocessing steps, and the fine-tuning process. Section 4 discusses the experimental setup, covering evaluation metrics, comparison methods, experimental results, and their analysis. The final section concludes the paper, summarizing our contributions and highlighting the significance of applying LLMs to the text to-SQL task, with a special emphasis on schema-less querying.

2. Related works

The task of text-to-SQL is to convert natural utterances into SQL queries. This field has attracted researchers in the NLP and the database community for decades.[4-9] The methodologies currently in use to handle this task can be broadly divided into three categories: rule-based methods, fine-tuning methods, and in-context learning (ICL) methods. Rule-based approaches, as highlighted in other studies,[7,10] utilize predefined templates to generate SQL queries. These methods show proficiency in certain scenarios but are limited by the necessity for manual rule formulation, which restricts their versatility across diverse domains.

Addressing the limitations of rule-based methods, recent research has ventured into more flexible approaches. The utilization of bi-directional long-short-term memory and convolutional networks[11] in Seq2Seq models has enhanced adaptability and effectiveness, though integrating structural database information remains a persistent challenge. Graph neural networks have emerged as a solution to this, with approaches that treat the database schema as a graph, as seen in other works.[12,13] Furthermore, the introduction of the MIMICSQL dataset and a model-based system by Translate-Edit Model for Question-to-SQL (TREQS) marked a significant advancement in text-to-SQL, particularly in the medical domain. Their model sets a robust baseline for subsequent evaluations. In our study, we use the MIMICSQL dataset and the TREQS model as benchmarks to evaluate and compare the effectiveness of our proposed method. In addition, fine-tuning pretrained language models like T5 have demonstrated improved performance in the text-to-SQL domain.[14-16] However, these fine-tuning methods typically require an extensive amount of labeled training data tailored to the specific task, and they are often susceptible to over-fitting. This limitation raises concerns about their versatility and efficiency in practical applications.

The advent of LLMs like GPT has opened new avenues in text-to-SQL tasks, particularly due to their ICL capabilities. These models often outperform fine-tuning methods in various NLP downstream tasks, especially in scenarios requiring few-shot or zero-shot learning. Nevertheless, the effectiveness of LLMs heavily relies on the design of input prompts, a factor that significantly influences the output quality.[17-19] The ICL performance of LLMs in text-to-SQL tasks, especially the impact of different prompts, has also been examined.

While basic prompting serves as a benchmark for assessing the fundamental capabilities of LLMs, more sophisticated prompt designs have shown to significantly enhance performance. Notably, a few-shot learning approach employing GPT-4 recently set a new benchmark in text-to-SQL tasks, achieving state-of-the-art results. However, this method necessitates manual input for demonstrations and tends to use a large number of tokens, requiring more time and resources.

This study extends the current advancements in LLMs within the Text-to SQL domain. Specifically, we fine-tune Flan-T5-based models on the MIMICSQL dataset. Each of these models is a sequence-to-sequence LLM that can be also used commercially. The model was published by Google researchers in late 2022 and has been fine-tuned on multiple tasks. It reframes various tasks into a text-to-text format, such as translation, linguistic acceptability, sentence similarity, and document summarization. Similarly, the architecture of the Flan-T5 model closely aligns with the encoder-decoder structure utilized in the original Transformer paper. The primary distinction lies in the size and nature of the training data; Flan-T5 was trained on an extensive 750 GB corpus of text known as the Colossal Clean Crawled Corpus (C4), and it comes with five variations: flan-t5-small (80M parameters, requiring 300 MB in memory), flan-t5-base (250M parameters, requiring 990 MB in memory), flan-t5-large (780M parameters, requiring 1 GB in memory), flan-t5-xl (3B parameters, requiring 12 GB in memory), and flan-t5-xxl (11B parameters, requiring 80 GB in memory). These models can be used for various NLP tasks out-of-the-box (with zero or few shot); however, to leverage its full potential and ensure optimal performance for specific applications, fine-tuning is a crucial step. Below are the main points and reasons highlighting the choice of fine-tuning FLAN-T5 for the specific text-to-SQL task:

(i) Fine-tuning FLAN-T5 is important to adapt the model to specific tasks and improve its performance on those tasks.
(ii) Fine-tuning allows for customization of the model to better suit the user's needs and data.
(iii) The ability to fine-tune FLAN-T5 on local workstations with CPUs makes it accessible to a wider range of users.
(iv) This accessibility is beneficial for smaller organizations or individual researchers who may not have access to GPU resources.
(v) Overall, fine-tuning FLAN-T5 is a valuable step in optimizing the model for specific use cases and maximizing its potential benefits.

Our emphasis on exploring schema-less approaches led us to investigate the viability and advantages of implementing text-to-SQL systems that depend less on explicit knowledge of database schema.

3. Data and methods

This section delineates the comprehensive methodology of our study, encompassing a detailed description of the dataset utilized, the architecture of the model employed, and the specifics of both the training and evaluation processes.

3.1. Dataset

The MIMICSQL dataset is a significant resource for question-to-SQL generation in the healthcare domain, comprising 10,000 question-SQL pairs. This large-scale dataset is based on the Medical Information Mart for Intensive Care III (MIMIC III) dataset, a widely used electronic medical records (EMR) database. It is divided into two subsets: one containing template questions (machine-generated) and the other featuring natural language questions (human-annotated).

3.1.1. Diversity and complexity of the dataset

The MIMICSQL dataset covers a wide range of patient information categories, including demographics, laboratory tests, diagnosis, procedures, and prescriptions. Those categories are embedded as a schema structure that outlines the database's tables, columns, and interrelationships, serving as a crucial guide for the models to comprehend the database structure and accurately formulate SQL queries. Table 1 illustrates the DEMOGRAPHIC table, while Table 2 presents the PROCEDURES table from the MIMICSQL dataset. This diversity reflects the complexity and multidimensionality of healthcare-related queries, as the SQL queries generated from these questions often involve multiple tables and columns.

3.1.2. Size and partitioning

The MIMICSQL dataset comprises approximately 10,000 examples, strategically partitioned into training and development (train and dev) sets, constituting 80% (8000 question-sql pairs), and a test set accounting for the remaining 20% (2000 question-sql pairs). This division facilitates both training and evaluation phases. Insights and statistical distributions from the MIMICSQL dataset are illustrated in Figure 1, and an illustrative example from the dataset is shown in Figure 2. Specifically, Figure 1A depicts the distribution of natural language questions, while Figure 1B focuses on the distribution of SQL query lengths. The presentation of natural language (NL) question and SQL query length distributions in the MIMICSQL dataset serves to reveal the dataset's inherent characteristics, aiding in the design of models capable of handling diverse language structures. In addition, it provides a basis for evaluating model performance by highlighting potential challenges associated with varying lengths. Understanding these distributions is crucial for both effective selections of the model training hyperparameters, especially the input and the output length, as well as assessment of the generalizability of the developed models to real-world applications.

Table 1. Example of the DEMOGRAPHIC table from theMIMICSQL database

SUBJECT_ID	HADM_ID	Gender	ADMISSION_TYPE	…
990	184231	F	EMERGENCY	…
17772	122127	M	NEWBORN	…
…	…	…	…	…
66411	178264	F	EMERGENCY	…

Table 2. EExample of the PROCEDURES table from the MIMICSQL database

SUBJECT_ID	HADM_ID	SHORT_TITLE	…
9258	183354	Procedure-one vessel	…
28588	141664	Insert endot. tube	…
…	…	…	…
66411	178264	Abdomen artery inc.	…

3.1.3. Challenges addressed

One of the key challenges addressed by the MIMICSQL dataset is the prevalence of abbreviations and potential typos in healthcare-related questions. This poses a significant obstacle to accurately generating the corresponding SQL queries, as the keywords provided in the questions may not precisely match those used in the EMR data. Consequently, the dataset presents a real-world scenario that requires models to effectively handle the nuances and complexities of healthcare-related queries.

3.2. Problem formulation

The SQL query generation task can be formulated as follows: Let D = {(Qi, SQL(Qi))} for i = 1, 2,…, N represents the dataset, where Qi represents the i-th natural language question, and SQL(Qi) is the corresponding ground-truth SQL query. The objective is to learn a mapping function F(Q; θ) parameterized by θ using the LLMs Flan-T5 Base and Flan-T5 Large:

$$SQL(Q)=F(Q;\theta) \qquad \text{(I)}$$

The training process involves minimizing the cross-entropy loss between the predicted SQL queries and the ground-truth queries:

$$Loss(\theta)= -\sum \log P\ (SQL(Q_i)|Q(_i;\theta) \qquad \text{(II)}$$

where $P(SQL(Q_i)|Q_i;\theta)$ denotes the probability of the model predicting the correct SQL query for the *i*-th natural language question (Q_i), given the model parameters θ. Our schema-less approach entails using only the input question Q as context during inference, without explicit database schema information.

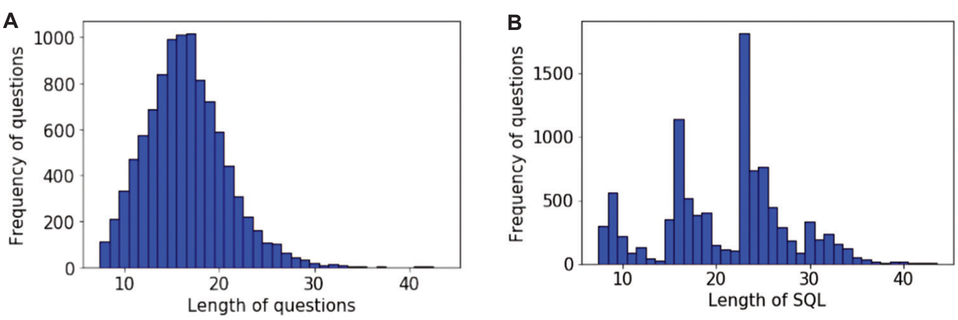


Figure 1. Distribution of the input and the output sequences in the MIMICSQL dataset. (A) The distribution of length on the questions (the input). (B) The distribution of length of SQL queries.

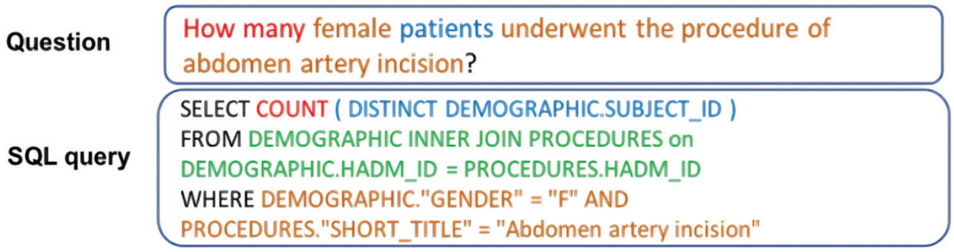


Figure 2. Illustrative example demonstrating how the MIMICSQL dataset utilizes the DEMOGRAPHICS and PROCEDURES tables to construct a response to a given question. This example employs color coding to distinctly indicate the correlations between components of the source question, the corresponding SQL query, and the SQL query template. Such examples highlight the dataset's structure and the complexity of mapping natural language questions to SQL queries.

3.3. Preprocessing

To effectively utilize the MIMICSQL dataset within the constraints of LLMs like Flan-T5, specific preprocessing steps are essential. These steps are designed to address the complexities of SQL syntax and the particular decoding capabilities of such models. A key consideration in this process is the model's vocabulary, as an excessive number of special tokens can detrimentally affect performance. The following outlines the detailed preprocessing steps undertaken:

3.3.1. Enclosing JSON objects in an array

Individual JSON objects in the MIMICSQL dataset were enclosed within an array to ensure a consistent JSON array structure. This step was essential for data manipulation and loading during subsequent processing.

3.3.2. Replacing SQL characters

Replacing symbols with their corresponding words in text preprocessing is a common practice in NLP and offers several advantages. One of the reasons for this practice is to address the absence of certain special characters, such as "<," "<=," and "<>," in the vocabulary of models like Flan-T5. On the other hand, this replacement enhances model understanding as words are typically more interpretable and easier for the model to learn. It can lead to improved generalization, as models have an easier time working with words, which are part of natural language, compared to arbitrary symbols. In addition, using words reduces ambiguity, as symbols can be context-dependent and unclear.

3.3.3. Converting JSON to CSV

After the preprocessing steps, the JSON data were transformed into CSV format to facilitate compatibility with data analysis and modeling libraries. The CSV format with "text" and "sql" columns allowed seamless data integration into the training and evaluation processes.

3.3.4. Data cleaning

The majority of the data cleaning process was conducted beforehand on the MIMIC III database, while building the MIMICSQL dataset, involving correcting errors in patient demographics, standardizing the format of clinical notes, and filtering out irrelevant data. On the MIMICSQL dataset, we only removed some duplicated statements and corrected some typos in the SQL queries.

3.3.5. Tokenization and text prefixing

In the preprocessing phase, we employed the Flan-T5 tokenizer to tokenize the input and the output texts. Flan-T5 utilizes a sub-word tokenizer to break down the input texts into smaller units, capturing both word-level and sub-word information. It is based on SentencePiece, a popular unsupervised text tokenizer and detokenizer, that employs a segmentation algorithm to divide the input texts into sub-word units, allowing the model to handle a wide range of vocabulary and linguistic nuances. By leveraging Flan-T5 Large's tokenization approach, we aim to capture the contextual information present in both complete words and sub-word units, enhancing the model's ability to comprehend and generate meaningful sequences during the subsequent stages of our text-to-SQL task. Moreover, we added a prefix "transform:" to each natural language question. The prefix is specific to the T5-based model, allowing it to recognize the text as a task to be transformed into SQL queries. For the target, a padding token ID is set to −100, an adjustment designed to disregard padding tokens during loss calculation.

Table 3 shows a running example for the preprocessing of two entries from the MIMICSQL dataset.

4. Experiments and results

In this section, we will describe the experimental setup, infrastructure details and evaluation metrics for the experiment, as well as a comparison with other models.

Table 3. A running example for the preprocessing steps

Original text	Preprocessed text	Tokenized text
Input: Get the number of patients who died in or before 2132 and were less than 72 years of age. Output: SELECT COUNT (DISTINCT DEMOGRAPHIC.'SUBJECT_ID') FROM DEMOGRAPHIC WHERE DEMOGRAPHIC.'AGE' < '72' AND DEMOGRAPHIC.'DOD_YEAR' <= '2132.0'	Input: transform: get the number of patients who died in or before 2132 and were less than 72 years of age. Output: select count (distinct demographic.'subject_id') from demographic where demographic.'age' less than '72' and demographic.'dod_year' less than or equal to '2132.0'	Input: ['▁transform', ':', '▁Get', '▁the', '▁number', '▁of', '▁patients', '▁who', '▁died', '▁in', '▁or', '▁before', '▁21', '32', '▁and', '▁were', '▁less', '▁than', '▁72', '▁years', '▁of', '▁age', '.'] Output: ['▁select', '▁count', '▁(', '▁distinct', '▁demographic', '.', "'", 'sub', 'ject', '_', 'i', 'd', "'", '▁', ')', '▁from', '▁demographic', '▁where', '▁demographic', '.', "'", 'age', "'", '▁less', '▁than', "'", '72', "'", '▁and', '▁demographic', '.', "'", 'd', 'o', 'd', '_', 'year', "'", '▁less', '▁than', '▁or', '▁equal', '▁to', '▁', "'", '2', '13', '2.0', "'"]
Input: calculate the minimum days for which patients aged 20 years or older were hospitalized. Output: SELECT MIN (DEMOGRAPHIC.'DAYS_STAY') FROM DEMOGRAPHIC WHERE DEMOGRAPHIC.'AGE' >= '20'	Input: transform: calculate the minimum days for which patients aged 20 years or older were hospitalized. Output: select min (demographic.'days_stay') from demographic where demographic.'age' greater than or equal to '20'	Input: ['▁transform', ':', '▁calculate', '▁the', '▁minimum', '▁days', '▁for', '▁which', '▁patients', '▁aged', '▁20', '▁years', '▁or', '▁older', '▁were', '▁hospital', 'ized', '.'] Output: ['▁select', '▁min', '▁(', '▁demographic', '.', "'", 'day', 's', '_', 'stay', "'", '▁', ')', '▁from', '▁demographic', '▁where', '▁demographic', '.', "'", 'age', "'", '▁greater', '▁than', '▁or', '▁equal', '▁to', '▁', "'", '20', "'"]

4.1. Experimental setup

A key objective of this study is to evaluate the practicality of training and inference processes. For this purpose, we employed a standalone machine equipped with a single Nvidia V100 GPU (16 GB vRAM) and 32 GB of system memory.

In our implementation, we utilized both Flan-T5 Base and Large versions of models, based on the original T5 encoder-decoder architecture, augmented with an instruction-finetune mechanism. This architecture consists of multiple layers of transformer blocks, including self-attention mechanisms and feed-forward neural networks. These transformer blocks enable the model to capture long-range dependencies and contextual information from input sequences.

The input and output sequence lengths were standardized to 1024 tokens. Sequences exceeding this length were truncated, while shorter sequences were padded using a pad token. This sequence length configuration enabled a maximum training batch size of two on our GPU setup. Various learning rates and optimizers were tested, ultimately leading to the selection of the Adam optimizer with a learning rate of 5e-5.

The fine-tuning phase for each model spanned five epochs, a decision based on empirical observations of convergence and generalization in preliminary trials. This epoch count offered an optimal balance between model performance and training duration, ensuring the models adequately learned from the MIMICSQL dataset patterns without overfitting. In terms of time, completing five epochs of training took approximately 4 h for the Flan-T5 Base version and 7 h for the Flan-T5 Large version. Table 4 provides in details the tuning scenario.

While evaluating other approaches, significant attention was given to the construction of prompts for LLaMA-2-7B, GPT-3.5-Turbo, GPT-4, and DeFog-SQLCoder to effectively generate SQL queries. For these models, prompts were meticulously designed to include schema information, facilitating the generation of accurate SQL queries. These prompts are essential for guiding the models through the task, leveraging their inherent language understanding capabilities. Detailed examples of these prompts are provided in Figure 3. Notably, our approach with the Flan-T5 models is distinct from this conventional method, removing the need for any schema information in the prompts. In this approach, the Flan-T5 models are fine-tuned in a way that the questions are the only inputs, and the database schema can be captured automatically in a more efficient fashion. This distinction underscores the uniqueness and efficiency of our methodology. It is important to mention that TREQS models, not being categorized as LLMs, did not necessitate such prompt-based approaches, but used the schema as input with the question, further differentiating our method from traditional practices.

4.2. Evaluation metric

The most common and used evaluation metric for the text-to-SQL task are logical form accuracy (LFA, or exact matching) and execution accuracy,[20] and since the MIMIQSQL community has not shared databases for computing the execution accuracy metric, the primary metric we used for evaluating the models performance in this study is LFA. In fact, it measures the percentage of exact string match between the generated SQL queries and the ground truth SQL queries. It is quantified as the percentage of instances where the model's predicted SQL

Table 4. Details of the tuning scenario

Parameter	Value	Description
output_dir	/logs_long	Directory where the trained model and logs will be saved
per_device_train_batch_size	1	Number of training samples per device (GPU) in each batch
per_device_eval_batch_size	1	Number of evaluation samples per device (GPU) in each batch
predict_with_generate	True	Whether to use generation during evaluation
fp16	False	Whether to use mixed precision training with FP16
learning_rate	5e-5	Learning rate for training
num_train_epochs	5	Number of training epochs
logging_dir	./logs_long	Directory for logging training metrics and logs
logging_strategy	steps	Strategy for logging training metrics (steps or epoch)
logging_steps	500	Interval for logging training metrics
evaluation_strategy	epoch	Strategy for evaluation during training (steps or epoch)
save_strategy	epoch	Strategy for saving checkpoints during training (steps or epoch)
save_total_limit	2	Maximum number of checkpoints to keep

```
Prompt for GPT3.5-turbo and GPT4:
"""You are a helpful assistant that generates SQL queries. \n
Given the schema {schema}, generate an SQL query to answer the following question:\n\n{question}\n\n
 SQL:"""

Prompt for Defog-SQLCoder:
"""### Task
Generate a SQL query to answer the following question and database schema:
`{question}`
`{schema}`"""

Prompt for LLAMA2-7B (finetuned on text2sql):
"""Below is an context that describes a sql query, paired with an question that provides further
information. Write an answer that appropriately completes the request.
### Context:
{schema}
### Question:
{question}
### Answer:"""

Prompt for our FlanT5 models, which includes just the question as input, prefixed with "transform":
"Transform: {question}"
```

Figure 3. Comparison of prompt construction for SQL query generation. This figure illustrates the detailed prompts used for LLaMA-2-7B, GPT-3.5-Turbo, GPT-4, and DeFog-SQLCoder, highlighting the inclusion of schema information in each, except our Flan-T5 models in compliant with the schema-less approach. Since TREQS is not regarded as an LLM, no prompt generated for that approach.

query exactly aligns with the ground-truth as described in formula (III).

$$LFA = \frac{NCLF}{TNI} \tag{III}$$

Where NCLF is the number of correct logical forms, and TNI is the total number of instances in the test set. A higher LFA score signifies a model's enhanced capability in accurately translating natural language questions into their corresponding SQL queries, reflecting a deeper understanding of the logic and relationships inherent in these representations. This makes LFA a particularly pertinent measure for the text-to-SQL task, as it directly gauges the fine-tuned models' effectiveness in interpreting natural language and generating precise SQL queries.

4.3. Results and discussion

The evaluation of various text-to-SQL models on the MIMICSQL test set has provided significant insights. The baseline TREQS model recorded an LFA of 0.48, which marginally increased to 0.55 with the incorporation of a recovery technique (TREQS + Recover). The current state-of-the-art model, Defog-SQLCoder, achieved an LFA of 0.65. In comparison, the LLMs GPT 3.5-Turbo and GPT-4 demonstrated robust performance with LFA scores of 0.60 and 0.70, respectively, highlighting their applicability. In addition, the LLaMA-2-7B model, which was fine-tuned for text-to-SQL tasks, attained an LFA of 0.60. Remarkably, our custom fine-tuned model, Flan-T5 Large, surpassed all these models with an LFA of 0.85.

Figure 4 presents a clear illustration of a sample natural language query, the ground truth SQL query that would accurately respond to this query, and the SQL queries generated by the LLMs used in our experiments, namely, LLaMA 2-7B, GPT-3.5-Turbo, GPT-4, and DeFog-SQLCoder, along with our Flan-T5 models. This comparison vividly highlights the differences in the query generation capabilities of each model, offering a tangible demonstration of their respective performances in the text-to-SQL context.

This outcome indicates that while existing models such as GPT-3.5-Turbo (20B parameters), LLaMA-2-7B (7B parameters), and Defog-SQLCoder (15B parameters) show commendable proficiency, our approach using the schema-less text-to-SQL with Flan-T5 Large, which has only 780M parameters, notably outperforms others. This demonstrates not only superior performance but also remarkable efficiency, offering transformative potential in both specific domains and broader applications. The detailed results are tabulated in Table 5.

The results from our comprehensive evaluation shed light on the text-to-SQL domain, underscoring the significance of language model-based models (LLMs) and the promising potential of schema-less approaches in healthcare. It is crucial to note that the LLMs under scrutiny, specifically LLAMA-2-7B and DeFog-SQLCoder, were fine-tuned on the text-to-SQL task, encompassing datasets such as MIMICSQL, thereby directly incorporating knowledge pertinent to this domain. On the other hand, the GPT models (GPT-3.5-Turbo and GPT-4) are renowned for their versatility in evaluating various NLP tasks, including text-to-SQL, due to their extensive pre-training on diverse corpora. While these models were not specifically fine-tuned on the MIMICSQL dataset, their broad exposure during pre-training to a wide array of textual and structured data may have contributed to their performance on the MIMICSQL test set. This factor is important to consider when interpreting the comparative performance of these

Figure 4. Sample SQL query generation. This figure illustrates a sample natural language query alongside the corresponding ground truth SQL query and the SQL queries generated by the evaluated LLMs (LLaMA-2-7B, GPT-3.5-Turbo, GPT-4, and DeFog-SQLCoder) and our Flan-T5 models. In addition, an augmented version of the ground truth query is presented, serving as an example of how we enriched the training data during the fine-tuning of our FlanT5 models. It is important to note that this augmentation was exclusively for training purposes; no data in the test set were altered or augmented in any manner.

models against our fine-tuned Flan-T5 models, which were explicitly trained on the MIMICSQL dataset.

Comparison of our fine-tuned Flan-T5 models with baseline TREQS, as well as other LLMs such as Defog-SQLCoder, LLaMA-2-7B, GPT-3.5, and GPT-4, highlights the significant impact of compact LLMs and illustrates the effectiveness of schema-less querying. A key finding is the ability of smaller models, such as Flan-T5, to generate SQL queries directly from natural language questions without relying on explicit database schema information. This approach challenges the traditional schema-dependent paradigm. Furthermore, as illustrated in the Table 6, our study emphasizes the importance of model competence. The Flan-T5 models demonstrated proficiency in interpreting complex medical terminology, even when certain terms were absent from the model's vocabulary.

Table 5. Comparison with current state-of-the-art models in the test set

Model	Parameter size	LFA
TREQS	2.8 M	0.48
TREQS+Recover	2.8 M	0.55
GPT-3.5-Turbo	20 B	0.60
LLAMA-2-7B	7 B	0.60
Defog-SQLCoder	15 B	0.65
GPT-4	-	0.70
Flan-T5 Base (Ours)	220 M	0.56
Flan-T5 Large (Ours)	**770 M**	**0.85**

Notes: Flan-T5 Large outperforms all other models. The parameter size of GPT-4 has not been disclosed by OpenAI. The boldface indicates the best model with the highest score achieved compared to other models, indicating its superior performance in the LFA metric.
Abbreviations: M: Million; B: Billion.

Table 6. Examples demonstrating proficiency in interpreting complex medical terminology using our fine-tuned Flan-T5 Large model

Input question	SQL query generated by Flan-T5 Large model
How many patients have diagnoses icd9 code 28803?	SELECT COUNT (DISTINCT DEMOGRAPHIC."SUBJECT_ID") FROM DEMOGRAPHIC INNER JOIN DIAGNOSES ON DEMOGRAPHIC.HADM_ID=DIAGNOSES.HADM_ID WHERE DIAGNOSES."ICD9_CODE" = '28803';
Which patients have a neolevo5 drug code?	SELECT COUNT (DISTINCT DEMOGRAPHIC."SUBJECT_ID") FROM DEMOGRAPHIC INNER JOIN PRESCRIPTIONS ON DEMOGRAPHIC.HADM_ID=PRESCRIPTIONS.HADM_ID WHERE PRESCRIPTIONS."FORMULARY_DRUG_CD" = 'NEOLEVO5';

In the first example, the natural language query is accurately transformed into SQL, selecting the count of distinct patient IDs from the DEMOGRAPHIC table where the corresponding HADM_ID in the DIAGNOSES table matches the specified ICD9 code "28803." This showcases Flan-T5's ability to understand and incorporate complex medical terminology, such as ICD9 codes, into SQL queries effectively. While in the second example, the model accurately generates an SQL query that selects the count of distinct patient IDs from the DEMOGRAPHIC table based on the presence of the specific drug code "NEOLEVO5" in the PRESCRIPTIONS table. Despite the unfamiliar term "neolevo5," Flan-T5 effectively incorporates it into the SQL query, showcasing its ability to comprehend and handle diverse medical and pharmaceutical terminology. This capability is vital in healthcare, where accuracy and domain-specific language are paramount. Our findings encourage further exploration into strategies that reduce model and data complexity, focusing on task-oriented models suitable for commodity hardware.

5. Conclusion

This study investigates the efficacy of smaller, task-specific language models compared to more complex LLMs in the Text-to-SQL task, with a focus on the healthcare domain using the MIMICSQL dataset. Our findings reveal the remarkable performance of the fine-tuned Flan-T5 models, particularly Flan-T5 Large, which achieved an LFA score of 0.85. This score surpasses the current state-of-the-art model, Defog-SQLCoder, as well as other advanced LLMs such as LLaMA-2-7B, GPT-3.5-Turbo, and GPT-4. Our approach, advocating for the removal of schema definitions from input prompts and training separate models for distinct schemas, has proven effective, requiring less hardware resources and data for training. These findings underscore the potential of tailored compact language models for domain-specific applications, opening avenues for more efficient and effective natural language understanding in specialized contexts.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare no competing interests.

Author contributions

Conceptualization: Youssef Mellah, Veysel Kocaman
Investigation: All authors

Methodology: Youssef Mellah, Veysel Kocaman
Formal analysis: Youssef Mellah, Veysel Kocaman, Hasham UI Haq
Writing – original draft: Youssef Mellah, Veysel Kocaman, Hasham UI Haq
Writing – review & editing: Veysel Kocaman, David Talby

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data used in this study can be found at: https://github.com/wangpinggl/TREQS/tree/master/mimicsql_data/mimicsql_natural_v2

References

1. Deng N, Chen Y, Zhang Y. Recent Advances in Text-to-SQL: A Survey of What We Have and What We Expect. In: *Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea. International Committee on Computational Linguistics*; 2022. p. 2166-2187.
2. Katsogiannis-Meimarakis G, Koutrika G. A survey on deep learning approaches for text-to-SQL. *VLDB J*. 2023;32:905-936.

   doi: 10.1007/s00778-022-00776-8
3. Wang P, Shi T, Reddy CK. Text-to-SQL Generation for Question Answering on Electronic Medical Records. In: *Proceedings of the Web Conference 2020* (*WWW '20*). New York, NY, USA: Association for Computing Machinery. p. 350-361.

   doi: 10.1145/3366423.3380120
4. Codd EF. A relational model of data for large shared data banks. *Commun ACM*. 1970;13(6):377-387.

   doi: 10.1145/362384.362685
5. Hemphill CT, Godfrey JJ, Doddington GR. The ATIS Spoken Language Systems Pilot Corpus. In: *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania*; 1990.
6. Dahl DA, Bates M, Brown M, *et al*. Expanding the Scope of the ATIS Task: The ATIS-3 Corpus. In: *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey*; 1994.
7. Zelle JM, Mooney RJ. Learning to Parse Database Queries Using Inductive Logic Programming. *In: Proceedings of the National Conference on Artificial Intelligence*; 1996. p. 1050-1055.
8. Popescu AM, Etzioni O, Kautz H. Towards a Theory of Natural Language Interfaces to Databases. In: *Proceedings of the 8th International Conference on Intelligent user Interfaces* (*IUI '03*). New York, NY, USA: Association for Computing Machinery; 2003. p. 149-157.

   doi: 10.1145/604045.604070
9. Bertomeu N, Uszkoreit H, Frank A, Krieger HU, Jörg B. Contextual Phenomena and Thematic Relations in Database QA Dialogues: Results from a Wizard-of-Oz experiment. In: *Proceedings of the Interactive Question Answering Workshop at HLT-NAACL*. New York, NY, USA: Association for Computational Linguistics; 2006. p. 1-8.
10. Saha D, Floratou A, Sankaranarayanan K, Minhas UF, Mittal AR, Özcan F. ATHENA: An ontology-driven system for natural language querying over relational data stores. *Proc VLDB Endow*. 2016;9(12):1209-1220.

    doi: 10.14778/2994509.2994536
11. Choi DH, Shin MC, Kim EG, Shin DR. RYANSQL: Recursively applying sketch-based slot fillings for complex text-to-SQL in cross-domain databases. *Comput Linguistics*. 2021;47(2):309-332.

    doi: 10.1162/coli_a_00403
12. Wang B, Shin R, Liu X, Polozov O, Richardson M. RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. United States: Association for Computational Linguistics; 2020. p. 7567-7578.

    doi: 10.18653/v1/2020.acl-main.677
13. Cao R, Chen L, Chen Z, Zhao Y, Zhu S, Yu K. LGESQL: Line Graph Enhanced Text-to-SQL Model with Mixed Local and Non-Local Relations. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Vol. 1 (Long Papers); 2021. p. 2541-2555.

    doi: 10.18653/v1/2021.acl-long.198
14. Raffel C, Shazeer N, Roberts A, *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res*. 2020;21(1):5485-5551.

    doi: 10.48550/arXiv.1910.10683
15. Scholak T, Schucher N, Bahdanau D. PICARD: Parsing Incrementally for Constrained Auto-Regressive Decoding from Language Models. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Punta Cana, Dominican Republic. Association for Computational Linguistics; 2021. p. 9895-9901.

    doi: 10.18653/v1/2021.emnlp-main.779 8111035487
16. Li H, Zhang J, Li C, Chen H. RESDSQL: Decoupling Schema Linking and Skeleton Parsing for Text-to-SQL. In: *Proceedings of the Thirty-Seventh AAAI Conference*

on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI' 23/IAAI' 23/EAAI '23), Vol. 37. Washington, DC, U.S: AAAI Press, 2023. p. 13067-13075.

doi: 10.1609/aaai.v37i11.26535

17. Min S, Lyu X, Holtzman A, *et al*. Rethinking the role of demonstrations: What makes in-context learning work? In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. p. 11048-11064.

    doi: 10.18653/v1/2022.emnlp-main.759

18. Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. What Makes Good In-Context Examples for GPT-3? In: *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 100-114.

    doi: 10.18653/v1/2022.deelio-1.10

19. Wei J, Wang X, Schuurmans D, *et al*. Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Info Process Syst*. 2022;35:24824-24837.

20. Yu T, Zhang R, Yang K, *et al*. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 3911-3921.

    doi: 10.18653/v1/D18-1425

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Development and analysis of medical instruction-tuning for Japanese large language models

Issey Sukeda[1]*, Masahiro Suzuki[2], Hiroki Sakaji[3], and Satoshi Kodera[1]

[1]Department of Cardiovascular Medicine, Graduate School of Medicine, The University of Tokyo, Bunkyo, Tokyo, Japan

[2]Department of Systems Innovation, School of Engineering, The University of Tokyo, Bunkyo, Tokyo, Japan

[3]Faculty of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido, Japan

Abstract

In the ongoing wave of impact driven by large language models (LLMs) like ChatGPT, the adaptation of LLMs to the medical domain has emerged as a crucial research frontier. Since mainstream LLMs tend to be designed for general-purpose applications, constructing a medical LLM through domain adaptation is a huge challenge. While instruction-tuning, particularly based on low-rank adaptation (LoRA), has become a frequently employed strategy to fine-tune LLMs recently, its precise roles in domain adaptation remain unknown. Here, we investigated how LoRA-based instruction-tuning improves the performance of Japanese medical question-answering tasks by employing a multifaceted evaluation of multiple-choice questions, including scoring based on "Exact match" and "Gestalt distance" in addition to the conventional accuracy. Our findings suggest that LoRA-based instruction-tuning can partially incorporate domain-specific knowledge into LLMs, with larger models demonstrating more pronounced effects. Furthermore, our results underscore the potential of adapting English-centric models for Japanese applications in domain adaptation, while also highlighting the persisting limitations of Japanese-centric models. This initiative represents a pioneering effort in enabling medical institutions to fine-tune and operate models without relying on external services.

Keywords: Medical large language models; Llama2; Instruction-tuning; Domain adaptation; Low-rank adaptation; QLoRA

***Corresponding author:**
Issey Sukeda
(sukeda-issei006@g.ecc.u-tokyo.ac.jp)

Citation: Sukeda I, Suzuki M, Sakaji H, Kodera S. Development and analysis of medical instruction-tuning for Japanese large language models.
Artif Intell Health. 2024;1(2): 107-116.
doi: 10.36922/aih.2695

Received: January 10, 2024

Accepted: March 13, 2024

Published Online: April 8, 2024

1. Introduction

The study and development of medical large language models (LLMs) like ChatGPT have the potential to revolutionize the field of medicine and healthcare in profound ways. These models, when fine-tuned and adapted to the medical domain, can assist healthcare professionals in numerous critical tasks, such as disease diagnosis, treatment planning, and patient care. Due to their vast language comprehension capabilities, LLMs may provide up-to-date information, suggest evidence-based treatment options, and even predict disease outcomes with a high degree of accuracy.

Domain adaptation remains a crucial approach for tailoring mainstream LLMs to the practical use in clinical environments, even after the surge of ChatGPT (https://chat.openai.com/), a powerful LLM service, that has revolutionized the way we interact with text and language by its astonishing ability to generate sentences. While these general-purpose models are powerful in zero-shot inference in unseen tasks, fine-tuned models may have the potential to outperform them in domain-specific tasks. Several works on domain adaptation within the medical field in the context of powerful English-centric LLMs[1-4] exist as well, but research in this direction is largely lacking in Japanese, highlighting the need to pioneer studies in non-English contexts. The drive to develop large-scale medical LLMs in one's native language is not only prevalent in Japan but also starting to mainstream in other non-English-speaking countries. In Japan, the sole precedent in the area of Japanese medical language model is the work of Sugimoto *et al.*,[5] who developed a Japanese medical language model named JMedRoBERTa based on RoBERTa, a BERT[6] -based model. This study is the first exploration along this line using large-scale GPT-models with a focus on text generation.

Moreover, ChatGPT utilization is impeded in clinical practices due to the concerns related to data privacy and security. The potential risks associated with data breaches or misuse of confidential patient information underscore the need for robust security measures and ethical considerations, further complicating its seamless integration into clinical settings. Hence, we need to consider domain adaptation using other LLMs for incorporating medical knowledge.

Recently, several parameter- efficient fine-tuning methods have been proposed, including low-rank adaptation (LoRA) and its quantized version (QLoRA),[7,8] where only the limited parameters are chosen as the target of the fine-tuning. Performed along with instruction-tuning, LoRA has demonstrated some success in acquiring conversational abilities and improving domain-specific performances such as financial question-answering tasks.[9,10] That being said, the ability and limitation of LoRA-based instruction-tuning have not been clarified in domain adaptation. "Superficial Alignment Hypotheses," which was proposed recently, provide a conjecture that fine-tuning does not contribute significantly to the acquisition of knowledge, but this topic remains controversial.[11] Therefore, we aim to investigate whether LoRA-based instruction tuning can be effective in acquiring domain-specific knowledge, especially medical knowledge.

The primary research questions guiding our study are as follows:

i. How and how much can domain knowledge be incorporated into LLMs by LoRA-based fine-tuning?
ii. Do larger English-centric LLMs outperform smaller Japanese-centric LLMs?
iii. Does the amount of fine-tuning hold significance?

To answer these questions, we conducted a comprehensive comparison between different LLMs fine-tuned with our own Japanese medical dataset by evaluating e ach model through medical question-answering approach. This enables us to clarify the strengths and limitations of incorporating domain-specific knowledge by LoRA, setting the stage for constructing enhanced versions of various domain-specific Japanese LLMs.

2. Related works

In recent years, there has been active research in constructing pretrained language models specialized for the medical domain. Before the emergence of GPT-3[12] in 2020 and ChatGPT in 2022, the prevailing trend in research involved building BERT[6] -based language models and evaluating them in classification tasks. In English-speaking regions, models such as BioBERT,[13] Med-BERT,[14] ClinicalBERT,[15] and PubMedBERT[16] have been proposed, leveraging medical literature databases such as PubMed and clinical records databases such as MIMIC-III.[17] Also in Japan, UTH-BERT[18] and JMedRoBERTa[5] have become available online. UTH-BERT[18] is the first medical pretrained language model in Japanese, pretrained by approximately 120 million lines of clinical texts. On the other hand, JMedRoBERTa[5] utilizes 11 million lines of journal articles in medicine, with the goal of accumulating information across a diverse range of content, encompassing basic research to case studies.

In the wake of GPT-3[12] and ChatGPT emergence, the focus of research shifted toward LLMs leveraging Transformer[19] accompanied with a steady increase in the parameter size of models. The primary tasks of interest in research also transitioned from classification tasks to medical text generation or medical question-answering. For the English-centric model, BioMedLM (formerly known as PubMedGPT),[20] BioGPT,[21] and BioMedGPT[22] have been proposed, harnessing the strength of the latest general-purpose LLMs. However, the currently available models have limited sizes: BioMedLM[20] has 2.7 billion parameters, BioGPT[21] is based on the GPT-2[23] architecture with 1.3 billion parameters, and BioMedGPT[22] comprises 10 billion parameters. On the other hand, Google has pursued its own path in developing medical models, including Med-PaLM[1] and Med-PaLM2[2] with 540 billion and 340 billion parameters, respectively; nonetheless, these models are not accessible to the public. To the best of our knowledge, there has been n research conducted to deepen the medical specialization of Japanese-centric model.

3. Data and methods

We conducted a comprehensive comparison between different LLMs fine-tuned with Japanese medical dataset, including those we have created ourselves. To determine whether one should start from a smaller Japanese model or a larger English model, we prepared OpenCALM-7B and Llama2-70B as base models. In addition, to observe the effectiveness of pretraining, we introduced a model additionally trained on medical documents. Subsequently, we applied medical instruction-tuning (LoRA, QLoRA) to each of them and evaluated performance based on the accuracy of medical question-answering tasks. The entire procedure is outlined in Figure 1. The models trained and used in our experiments are available at https://huggingface.co/AIgroup-CVM-utokyohospital.

3.1. Base model preparation

To create a Japanese-centric model, we utilized OpenCALM-7B (https://huggingface.co/cyberagent/open-calm-7b), an open-source Japanese foundation LLM with 6.5 billion parameters developed by CyberAgent, Inc. In addition, we trained a new base model MedCALM, which is based on OpenCALM-7B and continually pretrained on our own medical text dataset. Here, the training dataset consists of 2420 examples, and the evaluation dataset has 50 examples. The maximum token count is set to 768, and the batch size is set to 63. The model was trained for 2000 steps. On the other hand, we further used Llama2-70B-chat-hf (https://huggingface.co/meta-Llama/Llama-2-70b-chat-hf), a powerful English-centric LLM released by Meta Inc.[24] Hereinafter, it is referred to as Llama2-70B. The use of this model is governed by the Meta license (https://ai.meta.com/resources/models-and-libraries/llama-downloads/).

3.2. Medical instruction-tuning

Instruction-tuning refers to the process of fine-tuning or optimizing the behavior and output of the model by providing explicit instructions or guidance as a prompt during the generation of text.[25] We employed LoRA, one of the popular parameter-efficient fine-tuning methods provided in PEFT library,[7,26] since full fine-tuning, which retrains all model parameters, is unfeasible in our environment. LoRA freezes the pretrained model weights and inserts trainable rank decomposition matrices into each layer of the target model to reduce the number of trainable parameters for downstream tasks. Specifically, instead of directly updating the $d \times k$ parameter matrix of a linear layer in LLM from W_0 to $W_0+\Delta W$, LoRA updates a $d \times r$ matrix B and a $r \times k$ matrix A where BA is low-rank decomposition of ΔW, that is, $r \ll min\ (d,k)$.

Given our computational constraints, particularly the limited GPU memory, LoRA for OpenCALM-7B is feasible, but not for Llama2-70B. Instead, we opted for the quantized version, named QLoRA,[8] which is intended to trade off a slight performance drop for a significant reduction in model size, making the experiment using Llama2-70B feasible. Consequently, we applied LoRA to OpenCALM-7B and QLoRA to Llama2-70B, respectively. The hyperparameters of LoRA/QLoRA are listed in Table 1, which follow the default setting specified in PEFT library and QLoRA library, respectively.[8,26]

To perform medical instruction-tuning, we constructed a medical question-answer dataset containing 77422 records in instruction format. Initially, we reviewed two medical articles, one from the official journal of The Japanese Circulation Society (containing 3569 lines) and another from the Journal of the Japanese Society of Internal Medicine (JJSIM, containing 6120 lines), for input retrieval. Then, these texts were used as inputs for ChatGPT (gpt-3.5-turbo) to generate various question-answer pairs, resulting in 21365 records and 56057 records, respectively. Since ChatGPT is known to possess strong instruction-following ability, we utilized the following prompt template to construct instruction dataset with an overall good quality:

> ### Instructions: You are a machine designed to generate various question and answer pairs. Please create data with question (instruction) and answer (output) pairs based on the following input, considering it as prior knowledge. Format the data

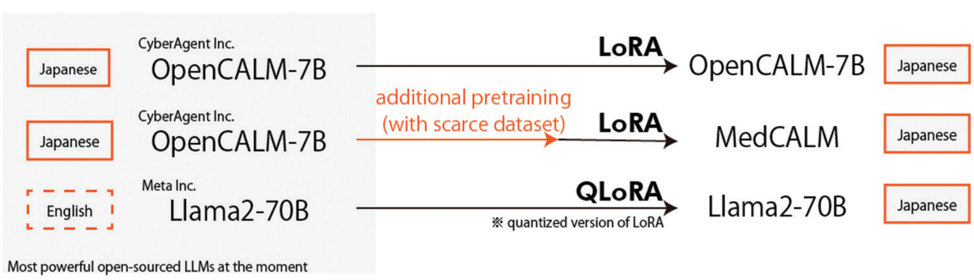


Figure 1. Overview of procedure of our medical instruction-tuning. Image created with Adobe Illustrator.

Table 1. LoRA/QLoRA parameters

	OpenCALM-7B	Llama2-70B
Fine-tuning method	LoRA	QLoRA
Learning rate	5e -5	2e-4
Input length	512	512
Target max length	512	512
Batch size	8	8
Fine-tuning steps	1k, 3k, 10k	0.9k, 3k
r of (Q) LoRA	8	64
α of (Q) LoRA	32	16
Dropout rate of (Q) LoRA	0.05	0.1
Target parameter	Query, Key, value	All linear layers

as "instruction": Question content, "output": Answer content, and do not include line breaks. Repeat this process 15 times and list one data pair per line.
Input: {input_text}

The number of epochs and steps was set to align with the overall computational time in each experiment. Using a larger model such as Llama2-70B increases the GPU memory usage per sample. To avoid this, memory usage can be reduced by decreasing the floating-point precision or by using gradient accumulation. In this study, we adopted 4-bit QLoRA on Llama2-70B. Since 4 bits is optimal in terms of the relationship between floating-point precision and model performance,[27] it is not desirable to reduce the floating-point precision any further. To experiment with less GPU memory, gradient accumulation was attempted by multiplying batch size calculation, for example, a batch size of 8 is calculated twice with four smaller mini-batch sizes. This approach allows for building larger models and reducing requirements for computing resources.

3.3. Evaluation by medical question-answering tasks

The state-of-the-art performance of English medical LLMs is typically evaluated using benchmark datasets such as MedQA (United States Medical Licensing Examination, USMLE),[28] MedMCQA,[29] and PubMedQA.[30] However, the availability of Japanese-curated medical task datasets is significantly limited, with IgakuQA (Japanese medical licensing exams)[31] being the only one available at present. Hence, in addition to IgakuQA, we prepared a new Q&A dataset JJSIMQA to assess the performance of each model in the medical domain. JJSIMQA is our own dataset comprising 5-choice questions included in JJSIM as appendices. Here are some samples from IgakuQA and JJSIMQA datasets:

An example from IgakuQA (originally in Japanese)
"problem_id": "116A1",
"problem_text": "Which of the following is incorrect regarding hypertension caused by obstructive sleep apnea?,"
"choices": {"a": "It often leads to nocturnal hypertension.", "b": "Weight reduction is recommended for obese patients.", "c": "Alpha-blockers are the first-line choice of medication.", "d": "Morning hypertension is frequently observed in home blood pressure measurements.", "e": "Continuous positive airway pressure (CPAP) therapy is expected to lower blood pressure."},
"text_only": True,
"answer": ["c"]
An example from JJSIMQA, 5-choice questions in JJSIM (originally in Japanese)
"problem_text": "Which of the following is incorrect about recent cases of hepatitis B in Japan? Choose one.",
"choices": {"a": "The HBs antigen positivity rate has significantly decreased due to the initiation of mother-to-child infection prevention programs.", "b": "HBV (hepatitis B virus) genotype Ae can become a carrier through horizontal transmission in adults.", "c": "In Japan, routine HBV vaccination began in October 2016.", "d": "HBV genotype C is more prevalent in the Tohoku and Miyako-Yaeyama regions.", "e": "Horizontal transmission of HBV during childhood is thought to be partly attributed to father-to-child transmission and communal living."},
"text_only": True,
"answer": ["d"]

The prompt template used for the evaluation follows the Alpaca-format,[32] where "problem_text" is incorporated in {instruction} and "choices" is incorporated in {input}:

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.
Instruction:
{Instruction}
Input:
{Input}
Response:

For evaluation in our experiments, these prompts were given in Japanese for OpenCALM-7B and in English for Llama2-70B. When generating the responses, we can specify parameters. In our experiments, *temperature* was set to 0.1, *max_new_tokens* to 256, *top_p* to 0.9, and *repetition_penalty* to 1.05. Question-answering samples that yielded null responses were excluded from the dataset.

Finally, we evaluated the output responses of each model by three different metrics: *Exact match*, *Gestalt score*, and *Accuracy*. While all these metrics aim to assess how effectively models can select the correct choice from five alternatives, they are defined with slight variations. Let R denote the response string and C^* denote the correct answer string among the five choices. *Exact match* takes the value of 1 if R and C^* exactly match at the string level, and 0 otherwise. *Gestalt score* is defined as the Gestalt distance between the response and the correct answer, which is calculated by a string matching algorithm that is based on the longest common subsequence: let K denote the longest matched string, then *Gestalt score* is calculated as $\text{GestaltScore}(R) = 2|K|/(|R|+|C^*|)$. Finally, *Accuracy* reflects the correctness by evaluating the choice closest to the model's response when measured using *Gestalt score*. Definitions are summarized as follows:

$S = \{C_1, C_2, C_3, C_4, C_5\}$: Choices,

C^* ($\in S$): The correct choice,

R: the response of the model,

$\text{ExactMatch}(R) = 1$ if $R = C^*$ else 0,

$\text{GestaltDistance}(R,C) = 2|K|/(|R|+|C|)$, K: the longest matched string between R and C,

$\text{GestaltScore}(R) = \text{GestaltDistance}(R,C^*)$,

$\text{Accuracy}(R) = 1$ if argmax_$\{C \in S\}$ $\text{GestaltDistance}(R,C) = C^*$ else 0.

All the evaluation metrics mentioned above take the value between 0 and 1, and the larger value indicates the better performance of the model.

3.4. Experimental settings

The whole dataset used in this work is summarized in Table 2. The experiments were run on 4 NVIDIA A100 with 80GB RAM each. All codes were implemented in Python, and the software and libraries we used include Transformers[33] and PEFT[26] from Hugging Face.

4. Results

4.1. The effect of medical instruction-tuning

The average score of experiments conducted for both 0-shot inference and 1-shot inference, measured by *Exact match*, *Gestalt score*, and *Accuracy* is summarized in Table 3 and Figure 2. The 0-shot inference refers to making responses without any specific examples, while the 1-shot inference refers to when one question-answer example is included in the input prompt. In Table 3, the top 2 scores in each row are highlighted in bold.

4.2. Comparison of our string-based evaluation metrics

Evaluation of LLMs is mainly conducted via manual evaluation[1] and automated evaluation based on rules. In automated evaluation methods, likelihood-based evaluation[34] is predominant. However, this evaluation method assesses the vectors outputted by the model rather than the actual generated strings, making it unsuitable for comparison with ChatGPT. To address this issue, our evaluation metrics are based on the strings actually outputted by the model. *Exact match* is a strict criterion where a response is considered correct only if it matches the correct answer precisely. Consequently, the number of correct answers is lower because even slight deviations are not considered correct. On the other hand, *Accuracy* is a relatively lenient metric where an output is considered correct as long as it is similar to the correct answer, even if it is not an exact match. This leads to a relatively higher number of correct answers as compared to *Exact match*, as deviations are tolerated to some extent.

Table 4 is a contingency table showing the number of question-and-answer (Q&A) samples where the model produced the correct answer. As a result, 112 question-answer samples are considered correct in terms of *Accuracy* but wrong in *Exact match*, whereas the reverse is not true. Among these 112 samples, many cases that were thought to be correct were not considered correct in the \textit{Exact match} evaluation. This was due to issues such as the model' s output being corrupted by token omissions in the tokenizer, or experiencing partial misrepresentation of Japanese characters, as observed in the examples listed in Table 5. This result implies that *Accuracy* is more suitable for evaluating performance in question-answering than *Exact match*, as it is more robust against the issues that models may potentially encounter. Further discussion in this regard is given in section 5.3.

4.3. Example responses from each model

We randomly created questions that ask each model the treatment of a symptom. This type of medical question is

Table 2. Datasets used in this work

Name	Source type	Format type	Purpose	Number of records
The Japanese circulation society	Academic journal	Alpaca format[32]	Instruction-tuning	21365
The *Journal of the Japanese Society of Internal Medicine*	Academic journal	Alpaca format[32]	Instruction-tuning	56057
IgakuQA[31]	Medical license exam	5-choice question	Evaluation	2002
JJSIMQA	Review questions	5-choice question	Evaluation	460

Table 3. Performance of Japanese medical question-answering tasks

	OpenCALM-7B				MedCALM				Llama2-70B		
Steps of QLoRA	**0**	**1k**	**3k**	**10k**	**0**	**1k**	**3k**	**10k**	**0**	**0.9k**	**3k**
Exact match (1s)	0	0.042	0.059	0	0.001	0	0	0	0.097	**0.200**	**0.173**
Gestalt score (1s)	0.053	0.186	0.087	0.078	0.028	0	0.002	0.035	0.247	**0.331**	**0.314**
Accuracy (1s)	0.177	0.190	0.148	0.174	0.164	0.150	0.150	0.165	0.200	**0.258**	**0.225**
Exact match (0s)	0	0.029	0.014	0.013	0	0.018	0.019	0.014	0.001	**0.180**	**0.169**
Gestalt score (0s)	0.033	0.114	0.141	0.120	0.032	0.096	0.116	0.085	0.071	**0.276**	**0.287**
Accuracy (0s)	0.170	0.182	0.166	0.193	0.185	0.172	0.240	0.183	0.170	**0.251**	**0.244**
Training hours	-	4.6	24	37	-	8.9	23.7	58.4	-	12.7	42.4

Notes: 0s and 1s denote 0-shot inference and 1-shot inference, respectively. The top 2 scores of each row are highlighted in bold. 0 steps denote the original base model.

Table 4. Number of Q&A samples where Llama2 (0.9k steps of QLoRA) produced the correct answer

	Correct in *Exact match*	Wrong in *Exact match*
Correct in *accuracy*	384	112
Wrong in *accuracy*	0	1425

not included in the instruction dataset nor the evaluation dataset. Table 6 shows the responses of each model to the following prompt, which was originally Japanese.

> ### Instruction:
> Please provide detailed instructions for the treatment to be administered to patients with the following diseases.
> ### Input:
> deep vein thrombosis
> ### Response:

Here, we observed that the original Llama2-70B generated English responses to some questions — 81% in 0-shot prompting and 15% in 1-shot prompting — while the other models responded completely in Japanese when prompt texts were given in Japanese.

5. Discussion

5.1. Numerical evaluation of the effects of fine-tuning

We observed notable score improvements with LoRA after an appropriate number of steps, particularly with Llama2-70B showing the most significant enhancement. This suggests that utilizing a more powerful English-centric model as the base model holds promise for domain adaptation even in Japanese contexts.

Regarding instruction-tuning, it has been controversial on whether, we should repeat epochs or just once. Our results showed that a single epoch (1k steps) of instruction-tuning improves the performance but increasing the number of epochs exacerbates the model. Furthermore,

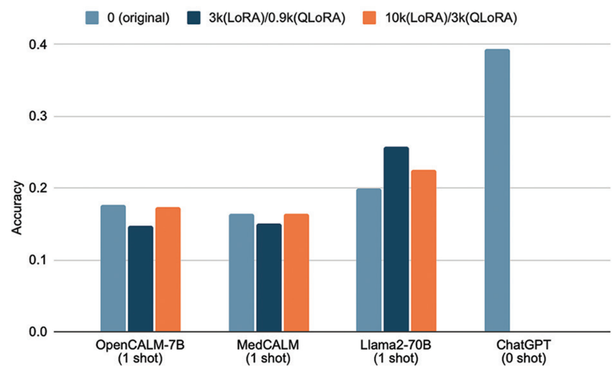


Figure 2. Comparison in accuracy of Japanese medical question-answering tasks. Image created with Google Spreadsheet.

additional pretraining did not contribute to performance improvement. Therefore, we conclude that conducting LoRA-based instruction-tuning for a single epoch without considering additional pretraining is a more practical and promising approach, especially when dealing with limited training data.

Note that in this study, we exclusively utilized medical documents closely related to the task for continual pretraining. However, we believe that the efficacy of additional pretraining could be further explored by incorporating a broader range of medical domain documents or by extracting and expanding from a general-purpose corpus. Determining the necessary amount of data for additional pretraining to improve performance in downstream tasks is a challenge, we will face in the future.

5.2. Deterioration of 1-shot performance

From Table 1, it is evident that every OpenCALM-based model except the original one experiences a decline in 1-shot inference scores rather than in 0-shot inference scores. This outcome highlights the fact that the original OpenCALM model clearly loses its capability to leverage example

Table 5. Some example outputs from Llama2-70B (0.9k steps of QLoRA)

Correct choice	Model's output	*Exact match*	*Gestalt score*	*Accuracy*
Workers' accident compensation insurance (労働者災害補償保険)	Similar but having tokenizing issue (労働者補償保険)	0	0.538	1
Functional menstrual disorders (機能性月経困難症)	Correct except third letter (機能的月経困難症)	0	0.736	1
Intra-articular hemorrhage (関節内出血)	Correct except the second letter (関箆内出血)	0	0.800	1
Monthly standard compensation for patient households (患者世帯の標準報酬月額)	Similar but having tokenizing issue (患者世帯患者世帯標準報酬月額)	0	0.666	1

Table 6. Example responses from each model

Model name	LoRA/ QLoRA	Response (exact output from LLM)
OpenCALM-7B	None	At the bedside, the nurse encouraged the patient to move their legs to assist in getting out of bed. The patient attempted to flex and extend their legs; however, they were unable to move the leg joints and needed assistance to stand up independently. The nurse approached the patient sitting on a bedside chair, provided shoulder support, lifted their arms, extended their elbows, and helped them rise slowly. The nurse then gently released their grip and returned to the chair. The nurse proceeded to call the next patient.*
	1k steps of LoRA	The procedure to be carried out for patients with the following disease involves the implementation of appropriate treatment and preventive measures.*
	3k steps of LoRA	Detailed treatment methods have not been provided.*
Llama2-70B	None	In some cases, a procedure called thrombol.
	0.9k steps of QLoRA	The treatment of deep vein thrombosis involves the use of thrombolytic agents and anticoagulants.*
	3k steps of QLoRA	The treatment for deep vein thrombosis includes pharmacological therapy to dissolve the blood clot as well as surgical interventions to remove the thrombus.*

Note: *Originally in Japanese.
Abbreviation: LLMs: Large language models.

responses provided within the context, whereas Llama2-70B retains this ability even after instruction-tuning.

5.3. Evaluation metrics

There have been some intensive arguments surrounding the evaluation of LLMs recently. Regarding the evaluation method of LLMs, there is still no unified "rule-of-thumb" method yet. While the existing metrics (*e.g.*, JGLUE[35]) or leaderboards (*e.g.*, Nejumi LLM leaderboard, http://wandb.me/nejumi) can assess the fluency of generated texts, they do not adequately evaluate the accuracy of domain-specific knowledge. It is noteworthy that three metrics used in our experiments also exhibit certain shortcomings. For example, *Exact match* cannot accurately score responses that, while conveying the correct meaning, do not match the text verbatim. *Gestalt score* is asymmetric and prone to multiple choices. Overall, our string-based metrics fall short in identifying phrases with different expressions but conveying the same meaning, and reflecting aspects such as fluency and medical accuracy. We argue that these features are not problematic in question-answering tasks where the model is required to output one or a few choices in short texts, but they become problematic when evaluating LLM for practical tasks, including medical report generation, where these aspects are crucial.

Furthermore, even the use of multiple-choice questions for evaluating LLMs has been controversial.[36,37] The development of even more superior evaluation metrics is eagerly anticipated.

5.4. Difficulty and limitations

While numerous LLM training techniques are still in the developmental stage, several shortcomings of training medical LLMs, like what we have done in this work, should be highlighted. First and foremost, the quantity and quality of data could be insufficient in our work. Preparing a medical dataset in instructional format can be expensive. In this study, we employed ChatGPT for automated generation, but this approach may become financially burdensome when preparing larger datasets. Data cleansing has also consistently posed challenges, and achieving perfect results in this work may not have been feasible.

Moreover, during the writing phase of this paper, Japanese LLMs that are considered to perform better than OpenCALM-7B, which was used in this study, have been released (see, *e.g.*, Rakuda benchmark, https://yuzuai.jp/benchmark). There is a possibility of obtaining different results when using them as the base model. Since one general implication suggested by the results of this experiment is that "a more powerful base model is preferable to start with," an overall performance improvement by upgrading the base model is highly expected.

6. Conclusion

In this paper, we explore the capabilities and limitations of LoRA through various comparative analyses in the medical domain. LoRA-based instruction-tuning, while avoiding an excessive number of steps, can partially integrate domain-specific knowledge into LLMs, with larger models demonstrating more pronounced effects. We also observe a decrease in performance after additional pretraining on scarce training dataset. Furthermore, our results underscore the potential of adapting larger English-centric models for Japanese applications in domain adaptation, while also highlighting the persisting limitations of Japanese-centric models including the deterioration of 1-shot performance after instruction-tuning. Our findings here suggest that, at present, the most promising approach in constructing a domain-specific LLM is applying QLoRA to larger English-centric base models.

Given the current situation, the clinical translation of medical LLMs into real-life applications still falls short of our expectations. To fully harness the potential of medical LLMs in healthcare settings, addressing both the performance limitations and the associated security and privacy concerns is imperative. Further research and development efforts are needed to enhance the accuracy and reliability of these models, ensuring they meet the rigorous standards required for clinical decision.

Furthermore, the integration of medical LLMs with other AI technologies, such as those utilized in electrocardiograms and electronic medical records, has the potential to amplify their impact significantly. By collaborating and cohesively using these AI systems along with medical LLMs, physicians can achieve a more comprehensive understanding of patient data, with which they could formulate more personalized treatment plans to improve patient outcomes.

Acknowledgments

None.

Funding

This study was supported by the Japan Agency for Medical Research and Development (Grant Number: JP23hk0102078h0003).

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Issey Sukeda, Satoshi Kodera
Formal analysis: Issey Sukeda
Investigation: Issey Sukeda, Satoshi Kodera
Methodology: Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji
Writing – original draft: Issey Sukeda
Writing – review & editing: Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Journal articles used in the study are available online in PDFs. ChatGPT is utilized for generating and cleansing the data. IgakuQA is available online. JJSIMQA is not made publicly available.

Further disclosure

Part of findings has been presented in *Deep Generative Models for Health* in NeurIPS 2023. In addition, a submission made to NeurIPS workshop is available on arXiv (https://doi.org/10.48550/arXiv.2310.10083).

References

1. Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. *Nature*. 2023;620:172-180.

   doi: 10.1038/s41586-023-06291-2
2. Singhal K, Tu T, Gottweis J, *et al*. Towards Expert-level Medical Question Answering with Large Language Models. *arXiv:2305.09617* [*arXiv Preprint*], 2023.

   doi: 10.48550/arXiv.2305.09617
3. Tu T, Azizi S, Driess D, *et al*. Towards generalist biomedical ai. *NEJM AI*. 2024;1(3).

   doi: 10.48550/arXiv.2307.14334
4. Wang G, Yang G, Du Z, Fan L, Li X. CLINICALGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation. *arXiv:2306.09968 [arXiv Preprint*], 2023.

   doi: 10.48550/arXiv.2306.09968
5. Sugimoto K, Iki T, Chida Y, Kanazawa T, Aizawa A. JMedRoBERTa: A Japanese Pre-trained Language Model

on Academic Articles in Medical Sciences (in Japanese). In: *Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing*; 2023.

6. Devlin J, Chang MW, Lee K, Toutanova K. 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.
7. Hu EJ, Wallis P, Allen-Zhu Z, *et al*. LoRA: Low-rank Adaptation of Large Language Models. In: *International Conference on Learning Representations*; 2021.
8. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems.* 2023;36:10088-10115.
9. Suzuki M, Hirano M, Sakaji H. From Base to Conversational: Japanese Instruction Dataset and Tuning Large Language Models. In: *2023 IEEE International Conference on Big Data (Big Data)*; 2023.
10. Xie Q, Han W, Zhang X, *et al*. PIXIU: A Comprehensive Benchmark, Instruction Dataset and Large Language Model for Finance. Advances in Neural Information Processing Systems. 2023;36:33469-33484.
11. Zhou C, Liu P, Xu P, *et al*. Lima: Less is More for Alignment. *Advances in Neural Information Processing Systems.* 2023;36:55006-55021.
12. Brown T, Mann B, Ryder N, *et al*. Language models are few-shot learners. *Adv Neural Inf Process Syst.* 2020; 33:1877-1901.
13. Lee J, Yoon W, Kim S, *et al*. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240.

    doi: 10.1093/bioinformatics/btz682
14. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med.* 2021;4:86.

    doi: 10.1038/s41746-021-00455-y
15. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv:1904.05342* [*arXiv Preprint*], 2019.

    doi: 10.48550/arXiv.1904.05342
16. Gu Y, Tinn R, Cheng H, *et al*. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc*. 2021;3(1):1-23.

    doi: 10.1145/3458754
17. Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data.* 2016;3:160035.

    doi: 10.1038/sdata.2016.35
18. Kawazoe Y, Shibata D, Shinohara E, Aramaki E, Ohe K. A clinical specific BERT developed using a huge Japanese clinical text corpus. *PLoS One*. 2021;16(11):e0259763.

    doi: 10.1371/journal.pone.0259763
19. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:5998-6008.
20. Bolton E, Hall D, Yasunaga M, Lee T, Manning C, Liang P. *Stanford CRFM Introduces PubMedGPT 2.7B*; 2022. Available from: https://hai.stanford.edu/news/stanford-crfm-introduces-pubmedgpt-27b [Last accessed on 2024 Apr 04].
21. Luo R, Sun L, Xia Y, *et al*. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform.* 2022;23:bbac409.

    doi: 10.1093/bib/bbac409
22. Luo Y, Zhang J, Fan S, *et al*. BioMedGPT: Open Multimodal Generative Pre-trained Transformer for Biomedicine. *arXiv:2308.09442* [*arXiv Preprint*], 2023.

    doi: 10.48550/arXiv.2308.09442
23. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI Blog*. 2019;1:9.
24. Touvron H, Martin L, Stone K, *et al*. Llama 2: Open Foundation and Fine-tuned Chat Models. *arXiv:2307.09288* [*arXiv Preprint*], 2023.

    doi: 10.48550/arXiv.2307.09288
25. Wei J, Bosma M, Zhao V, *et al*. Fine-tuned Language Models are Zero-shot Learners. In: *International Conference on Learning Representations*; 2022.
26. Mangrulkar S, Gugger S, Debut L, Belkada Y, Paul S. *PEFT: State-of-the-art Parameter-Efficient Fine-tuning Methods*; 2022. Available from: https://github.com/huggingface/peft [Last accessed on 2024 Apr 04].
27. Dettmers T, Zettlemoyer L. The Case for 4-bit Precision: K-bit Inference Scaling Laws. In: *International Conference on Machine Learning*. PMLR; 2023.
28. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl Sci*. 2021;11(14):6421.

    doi: 10.3390/app11146421
29. Pal A, Umapathi LK, Sankarasubbu, M. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical Domain Question Answering. In: *Proceedings of the Conference on Health, Inference, and Learning (2022)*; 2022. p. 248-260.
30. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: A Dataset for Biomedical Research Question Answering. In:

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019. p. 2567–2577.

doi: 10.18653/v1/D19-1259

31. Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev D. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. *arXiv:2303.18027* [*arXiv Preprint*], 2023.

    doi: 10.48550/arXiv.2303.18027

32. Taori R, Gulrajani I, Zhang T, *et al*. *Stanford Alpaca: An Instruction-following Llama Model*; 2023. Available from: https://github.com/tatsu-lab/stanford_alpaca [Last accessed on 2024 Apr 04].

33. Wolf T, Debut L, Sanh V, *et al*. Transformers: State-of-the-Art Natural Language Processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*; 2020. p. 38-45.

34. Gao L, Tow J, Biderman S, *et al*. A framework for few-shot language model evaluation. *Zenodo*. 2023;v0.0.1.

    doi: 10.5281/zenodo.5371629

35. Kurihara K, Kawahara D, Shibata T. JGLUE: Japanese General Language Understanding Evaluation. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*; 2022. p. 2957-2966.

36. Pezeshkpour P, Hruschka E. Large Language Models Sensitivity to the Order of Options in Multiple-choice Questions. *arXiv:2308.11483* [*arXiv Preprint*], 2023.

    doi: 10.48550/arXiv.2308.11483

37. Zheng C, Zhou H, Meng F, Zhou J, Huang M. Large Language Models are not Robust Multiple Choice Selectors. *arXiv:2309.03882* [*arXiv Preprint*], 2023.

    doi: 10.48550/arXiv.2309.03882

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Factors associated with social determinants of health mentions in PubMed clinical case reports from 1975 to 2022: A natural language processing analysis

Julio Bonis*, Veysel Kocaman, and David Talby

John Snow Labs Inc., Delaware, United States of America

Abstract

Social determinants of health (SDoH) significantly influence health outcomes, accounting for nearly 40% of such outcomes globally. These determinants, pivotal in understanding health disparities, are insufficiently documented in clinical settings and academic clinical narratives. To address this gap, we examined clinical case reports from PubMed (1975–2022) to identify mentions of six specific SDoH, employing a pre-trained named-entity recognition (NER) model from Spark natural language processing (NLP). Multivariate logistic regression was utilized to investigate associations between article characteristics and the documentation of SDoH. From 463,546 reports, 4.4% mentioned SDoH, with race/ethnicity being the most dominant mention. Race/ethnicity was often cited by sub-Saharan African authors (adjusted odds ratio [AOR]: 4.47) and in general medicine (AOR: 2.18). Marital status mentions appeared predominantly in psychiatry (AOR: 2.60) and gynecology (AOR: 2.47). Sexual orientation mentions were correlated with infectious diseases (AOR: 25.00) and varied by authorship regions, with stronger associations observed in South America (AOR: 4.04) and North America (AOR: 2.15), and comparatively weaker associations noted in the Indian subcontinent and the Middle East (AOR: 0.16). Immigrant status mentions were closely related to infectious diseases (AOR: 4.51), gynecology (AOR: 4.25), and certain geographies. Homelessness mentions were more prominent in forensic medicine (AOR: 14.92) and in both infections (AOR: 6.36) and mental disorders (AOR: 5.80). Spiritual belief mentions were more prominent with sub-Saharan authors (AOR: 9.17) and psychiatry (AOR: 7.61). SDoH mentions in medical literature were also determined by the diagnosis, cultural background, and journal type. The limited SDoH registration emphasized their overlooked significance. Disproportionate emphasis on specific relationships, such as sexual orientation with infectious diseases, can perpetuate biases and stereotypes. Innovative tools such as Spark NLP offer promise in advancing research using electronic health records (EHRs), but a standardized approach to SDoH reporting and vigilant AI training is crucial for unbiased health-care analysis.

Keywords: Social determinants of health; Natural language processing; Clinical case reports; Ethnicity; Marital status; Immigrant status; Homeless; Spiritual beliefs

***Corresponding author:**
Julio Bonis
(julio@johnsnowlabs.com)

Citation: Bonis J, Kocaman V, Talby D. Factors associated with social determinants of health mentions in PubMed clinical case reports from 1975 to 2022: A natural language processing analysis. *Artif Intell Health.* 2024;1(2): 117-131. doi: 10.36922/aih.2737

Received: January 14, 2024

Accepted: March 18, 2024

Published Online: April 17, 2024

1. Introduction

Social determinants of health (SDoH) are fundamental conditions that underpin the health disparities experienced by individuals globally. These determinants are the circumstances in which people are born, grow, work, and live, and they encompass factors such as socioeconomic status, housing, food security, and exposure to violence or stress.[1,2] Notably, these conditions have been proven to shape health outcomes to such an extent that up to 40% of health outcomes are attributed to SDoH challenges.[3,4]

Significantly, SDoH not only impacts health outcomes but also has discernible effects on health-care utilization. For instance, unmet social needs, a facet of SDoH, have been tied to clinical outcomes such as uncontrolled diabetes,[5] hypertension,[6] and increased hospital readmissions for heart failure.[7] There is also evidence suggesting that moving from a high-poverty neighborhood to one with lower poverty levels can lead to reductions in conditions such as extreme obesity and diabetes, emphasizing the role of environmental factors on health.[8]

Given the undeniable influence of SDoH on health, there have been initiatives to incorporate SDoH screening into health-care delivery, with proposals to standardize the methods for capturing this information in electronic health records (EHRs).[9] Advocates believe that documenting SDoH systematically at the point of care would bolster the identification of patients' risk factors and streamline referrals to social services, fostering a more holistic approach to patient care.[10,11]

However, the current reality paints a different picture. Despite the evident significance of SDoH, they remain underrepresented in clinical documentation. Recent studies have indicated that a mere 2% of patients visiting community health centers had at least one documented SDoH,[12] a figure that was confirmed by the analysis of the ICD10 codes in different studies.[13,14] Moreover, another study examining over a million unique patient EHRs found that only a small percentage contained mentions of social isolation, housing issues, or financial strain,[15] a finding that has been replicated in other studies.[16] However, other analyses conducted in the primary care context have reported slightly higher proportions i.e., 7% of patients with SDoH documented in Spain[17] and 4% to 18% in the United States (US).[18] These findings indicate that utilizing EHRs for SDoH documentation is insufficient, and a systemic approach involving education, policy redesign, and incentives might be necessary to boost documentation.[9]

These findings are concerning as a discrepancy in SDoH documentation could be indicative of a broader oversight in clinical decision-making. Within the domain of medical literature, clinical case reports serve as a reflection of the priorities and perspectives of health-care professionals. The choices they make in detailing specific patient information — what they choose to include or exclude—offer insights into what they deem significant or irrelevant. As such, the inclusion or omission of SDoH in these published reports can act as a barometer of their importance within the health-care community. By analyzing the frequency and context of SDoH mentions in these clinical cases, one can gauge the weight and significance attributed to these factors by health-care professionals when communicating notable clinical findings to a wider scientific audience.

Natural language processing (NLP) has become an indispensable tool in the medical domain, revolutionizing the extraction and analysis of complex data from clinical texts and patient records. Recent publications[19,20] highlight the crucial role of NLP in identifying, categorizing, and analyzing health-related information from unstructured content as clinical narratives. The advancements in NLP technologies, such as context-aware models like Bidirectional Encoder Representations from Transformers (BERT)[21] and BioBERT,[22] have dramatically enhanced our ability to process vast datasets, thereby transforming traditional health-care data analysis methods.[23-26] These innovations offer deeper insights into the prevalence and impact of SDoH, previously obscured in clinical documentation.[27] For instance, research has demonstrated that NLP-based systems can identify clinical events with significantly higher precision and sensitivity compared to traditional methods. One study demonstrated that an NLP system identified approximately four times as many clinical events as standard approaches, with a positive predictive value (PPV) of 74%, a stark improvement over the 31% PPV of methods relying solely on diagnostic codes.[28] In another study, the precision of selected cases increased from 46% to 86% after incorporating NLP methods that followed structured-based case selection with a sensitivity of 77%.[29] These examples highlight the transformative impact of NLP in enhancing the detection and characterization of SDoH and clinical events from medical narratives, enabling a more nuanced and comprehensive analysis of health-care data.

Our study utilizes advanced NLP technology to meet the need for improved documentation and understanding of SDoH in clinical settings. We investigated factors influencing the mention of SDoH in publicly available clinical case reports and how this knowledge could inform the development of more effective policies for SDoH reporting. In addition, our analysis identified potential stereotypes or discrimination in artificial intelligence (AI) models trained in the medical literature. We believe that our research adds to the discussion on SDoH, which could consequently enhance AI tools and policies for unbiased reporting of these determinants.

2. Methods

We obtained the latest annual PubMed baseline (available on September 1, 2023) through File Transfer Protocol (FTP) and parsed the search results to exclusively display publications tagged as "Clinical Case Report," yielding a total of 1,643,513 reports. We refined the search for articles published from January 1, 1975, to December 31, 2022. In addition, we employed a set of regular expressions to only include papers with abstracts that present a genuine clinical narrative about individual patients, rather than reports of aggregated case series. These were designed to pinpoint abstracts that mention both the age and gender of a single patient, resulting in the identification of 463,546 relevant articles (Figure 1).

To delineate the content of each article, we utilized a deep learning-based sentence boundary detection model[25,30] and produced a list of sentences for every article. Our focus was strictly on sentences that mentioned the patients' age and gender and identified using the same set of regular expressions. These sentences were then input into a pre-trained named-entity recognition (NER) model from John Snow Labs (JSL), designed to identify mentions associated with various SDoH and based on a proprietary fine-tuned BERT architecture.[31,32]

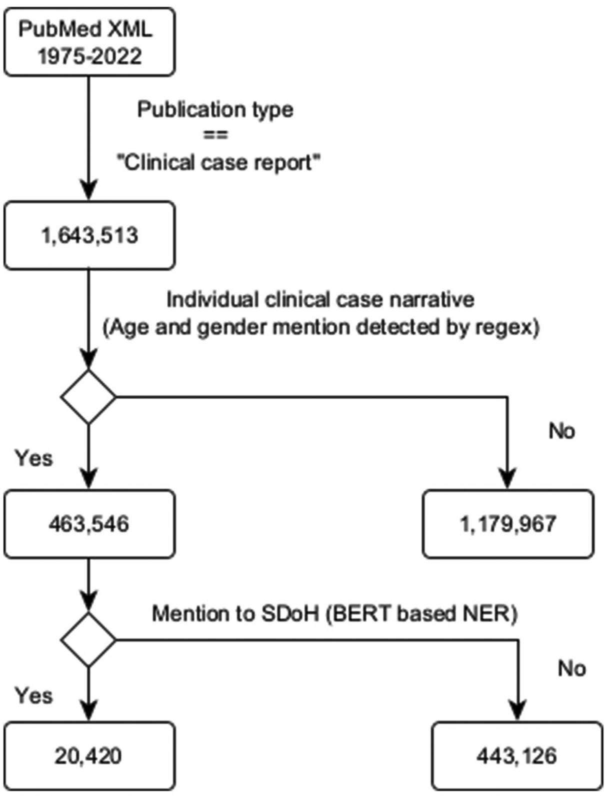


Figure 1. Workflow diagram illustrating the selection process of clinical case reports. The figure was created with yEd.
Abbreviations: BERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; NER: Named-entity recognition; SDoH: Social determinants of health; XML: Extensible markup language.

The accuracy of the model was assessed with an external dataset from JSL, encompassing 9,743 sentences and 198,698 tokens with manually annotated mentions to SDoH, namely race/ethnicity (n = 72), sexual orientation (n = 20), marital status (n = 193), housing (n = 371), population subgroup (n = 19), and spiritual beliefs (n = 90). This external test also compared the outcomes to generative pre-trained transformer (GPT)-3.5[33] and GPT-4.[34] In addition, an internal validation reviewed the precision for each SDoH entity found by the model in the PubMed dataset used in this study.

Besides the formal evaluation that considered the specific assertions of entities, our internal analysis prioritized identifying factors linked to SDoH mentions in clinical narratives. Hence, it was unnecessary to delve into the precise details or assertions regarding SDoH, such as a patient's marital status, whether they were married, unmarried, or if their marital status was unspecified. Our main interest was determining whether any SDoH mention, like marital status, was made, irrespective of its actual status or value. This method streamlined the extraction process by removing the need to navigate the intricacies associated with each SDoH status.

Consequently, our approach aligned with the study's objective to simply ascertain the occurrence of SDoH mentions within clinical documentation. Age and gender, used as selection criteria, were omitted from the SDoH evaluation. We targeted six specific SDoH, i.e., race/ethnicity, marital status, population group/immigrant status, sexual orientation, spiritual beliefs, and housing/homelessness, and analyzed them based on recall, precision, exclusion of individual behavior determinants not essentially social, and minimum corpus occurrence of 50 matches.

The journals' geographic origins were identified from PubMed records, and the first author's geographic origin was obtained from their reported affiliation. The main diagnosis was obtained from PubMed's Medical Subject Headings (MeSH) codes corresponding to disease or mental condition categories. Only root primary disease categories (e.g., respiratory tract, neurological, and mental conditions) were used during the analysis.

To analyze the link between article features and SDoH mentions, we conducted six logistic regression analyses using the Python package statsmodels 0.14.0 to gauge the adjusted odds ratio (AOR) for each article trait. We also employed a stepwise additive method,[35] where features that could enhance the likelihood of the model were sequentially incorporated with a P-value threshold of 0.001 for the likelihood ratio test.

3. Results

3.1. Study population and data inclusion

We analyzed a comprehensive dataset comprising 463,546 clinical case reports indexed in Medline from 1975 through 2022. The distribution of the articles based on four key information (author's geographic region, journal's geographic region, journal specialty, and clinical diagnosis) is displayed in Table 1.

3.2. Recall and precision of identifying mentions of the social determinants of health

In our corpus analysis, the SDoH identification precisions were 99.3% (95% confidence interval [CI]: 99.2 – 99.4%) for race/ethnicity, 90.2% (95% CI: 88.8 – 91.4%) for marital status, 90.8% (95% CI: 86.9–93.6%) for population group, 97.4% (95% CI: 95.6 – 98.4%) for sexual orientation, 100% (95% CI: 94.6 – 100%) for housing, and 98.4% (95% CI: 91.7 – 99.7%) for spiritual beliefs.

During external validation, the precision results were 97.4% (95% CI: 86.5 – 99.5%) for race/ethnicity, 100% (95% CI: 92.3 – 100%) for marital status, 88.9% (95% CI: 56.5 – 98.0%) for population group, 93.8% (95% CI: 71.7 – 98.9%) for sexual orientation, 98.6% (95% CI: 92.3 – 99.7%) for housing, and 83.0% (95% CI: 70.8 – 90.8%) for spiritual beliefs.

The recalls in the external validation were 90.2% (95% CI: 77.5 – 96.1%) for race/ethnicity, 97.9% (95% CI: 88.9 – 99.6%) for marital status, 88.9% (95% CI: 56.5 – 98.0%) for population group, 100% (95% CI: 79.6 – 100%) for sexual orientation, 85.2% (95% CI: 75.9 – 91.37%) for housing, and 83.0% (95% CI: 70.8 – 90.8%) for spiritual beliefs.

In our analysis comparing the recall and precision of the JSL SDoH-NER model with those of zero-shot learning (i.e., GPT-3.5 and GPT-4), both JSL and GPT-4 displayed comparable results. Notably, some differences were evident: JSL outperformed GPT-4 in precision for marital status (p = 0.005; GPT-4 scored 82.9%; 95% CI: 67.3–91.9%) and housing (p < 0.001; GPT-4 scored 82.9%; 95% CI: 67.3–91.9%). The results of this comparison are detailed in Figures S1 and S2.

3.3. Prevalence of social determinants of health mentions

Among the total case reports examined, 20,420 (4.4%) case reports included references to at least one SDoH category. A breakdown revealed that 17,765 case reports specifically mentioned race/ethnicity, followed by 1,991 articles that discussed marital status, 524 on sexual orientation, 284 on immigrant status, 63 on spiritual beliefs, and 60 on homelessness. The mean and confidence intervals of the mentioned rates within the study period are summarized in Table 2.

The analysis of the proportion of clinical cases reporting SDoH within the study period indicated a statistically significant association between publication year and race/ethnicity (P < 0.001), sexual orientation (P < 0.001), and homelessness (P < 0.001), respectively. Notably, there was a peak of sexual orientation mentions from 1980 to 1995, and we hypothesized that this could be related to the rise of acquired immunodeficiency syndrome (AIDS) cases, as depicted in Figure S3. There was also a prominent increase in race/ethnicity mentions between 2011 and 2013 (Figure S4) and a less evident but statistically significant increase in homelessness mentions since 1990.

3.4. Factors associated with reporting social determinants of health

3.4.1. Race/ethnicity

Significant associations were observed between the author's geographic origins and the frequency of race/ethnicity mentions. Authors from sub-Saharan Africa were most likely to discuss race/ethnicity (AOR: 4.47; 95% CI: 3.96 – 5.04), followed by the Caribbean (AOR: 3.31; 95% CI: 2.24 – 4.89), Southeast Asia (AOR: 2.89; 95% CI: 2.58 – 3.25), East Asia (AOR: 2.00; 95% CI: 1.90 – 2.09), and North America (AOR: 1.77; 95% CI: 1.68 – 1.86). Conversely, authors from the Indian subcontinent (AOR: 0.69; 95% CI: 0.62 – 0.76) and Middle East (AOR: 0.77; 95% CI: 0.70 – 0.84) were less inclined to mention race/ethnicity in their case reports.

The journal's geographic region also exerted an independent influence on race/ethnicity mentions. Journals originating from Australia-Oceania (AOR: 1.34; 95% CI: 1.17 – 1.53) and Western Europe (AOR: 1.30; 95% CI: 1.18 – 1.43) were slightly more prone to include race/ethnicity. In contrast, journals from East Asia (AOR: 0.48; 95% CI: 0.43 – 0.54), Eastern Europe (AOR: 0.54; 95% CI: 0.45 – 0.64), and South America (AOR: 0.55; 95% CI: 0.43 – 0.69) had much fewer race/ethnicity mentions than expected.

Table 1. Information on the analyzed articles (*n*=463546)

Information	Distribution	Number of articles	Percentage distribution of articles (%)
Author's geographic region	Known	334666	72.20
	East Asia	95527	28.54
	Western Europe	94950	28.37
	North America	72892	21.78
	Middle East	23631	7.06
	Indian subcontinent	13809	4.13
	Eastern Europe	9079	2.71
	South America	8299	2.48
	Australia and Oceania	6283	1.88
	Southeast Asia	3383	1.01
	Sub-Saharan Africa	2688	0.80
	North Africa	2440	0.73
	Central America	1395	0.42
	Caribbean	265	0.08
	Central Asia	25	0.01
	Unknown	128880	27.80
Journal's geographic region	Known	462600	99.80
	Western Europe	196878	42.56
	North America	150489	32.53
	East Asia	72101	15.59
	Eastern Europe	11157	2.41
	Australia and Oceania	8674	1.88
	Indian subcontinent	6780	1.47
	Middle East	6470	1.40
	South America	3759	0.81
	Sub-Saharan Africa	3101	0.67
	Southeast Asia	1657	0.36
	North Africa	617	0.13
	Central America	612	0.13
	Caribbean	305	0.07
	Unknown	946	0.20
Journal specialty	Known	423452	91.35
	General medicine	85521	20.20
	Surgery	77849	18.38
	Neurology	30533	7.21
	Oncology	23319	5.51
	Pediatrics	19518	4.61
	Cardiology	19393	4.58
	Dermatology	17516	4.14
	Pathology	17205	4.06
	Ophthalmology	15254	3.60
	Gastroenterology	12554	2.96
	Laboratory	12123	2.86

(*Cont'd...*)

Table 1. (*Continued*)

Information	Distribution	Number of articles	Percentage distribution of articles (%)
	Radiology	11792	2.78
	Urology	11641	2.75
	Gynecology	9354	2.21
	Infectiology	9190	2.17
	Traumatology	8028	1.90
	Hematology	6968	1.65
	Anesthesiology	6398	1.51
	Endocrinology	4996	1.18
	Neurology	4911	1.16
	Rheumatology	3687	0.87
	Nephrology	3555	0.84
	Psychiatry	3273	0.77
	Dentistry	2655	0.63
	Forensic	2332	0.55
	Public Health	1165	0.28
	Rehabilitation	1140	0.27
	Genetics	931	0.22
	Allergy	651	0.15
	Unknown	40094	8.65
Diagnosis	Neoplasms	154185	33.26
	Pathological signs and symptoms	117438	25.33
	Nervous system diseases	83899	18.10
	Infections	68717	14.82
	Cardiovascular diseases	67711	14.61
	Digestive system diseases	40355	8.71
	Musculoskeletal diseases	38527	8.31
	Urogenital diseases	37470	8.08
	Respiratory tract diseases	31740	6.85
	Hemic and lymphatic diseases	30350	6.55
	Skin and connective tissue diseases	22786	4.92
	Nutritional and metabolic diseases	20015	4.32
	Wounds and injuries	19674	4.24
	Eye diseases	19475	4.20
	Congenital, hereditary, and neonatal diseases	15903	3.43
	Stomatognathic diseases	9776	2.11
	Endocrine system diseases	9768	2.11
	Mental disorders	9109	1.97
	Chemically-induced disorders	7722	1.67
	Immune system diseases	7054	1.52
	Otorhinolaryngologic diseases	4339	0.94
	Occupational diseases	914	0.20
	Animal diseases	394	0.08
	Disorders of environmental origin	2	0.00

Note: Percentages of known characteristics are expressed relative to the total number of known articles; the cumulative percentage of diagnoses is more than 100% as a single article can have one or more assigned diagnoses; the list of diagnoses is based on the Medical Subject Headings (MeSH).

Table 2. Average SDoH mentions from clinical case reports (*n*=463546) between 1975 and 2022

SDoH	SDoH mentions (95% CI)
Race/ethnicity	383.24 (377.71–388.77)
Marital status	42.95 (41.06–44.83)
Sexual orientation	11.30 (10.34–12.27)
Immigrant status	6.13 (5.41–6.84)
Spiritual beliefs	1.36 (1.02–1.69)
Homelessness	1.29 (0.97–1.62)

Abbreviations: CI: Confidence interval; SDoH: Social determinants of health.

The specialty of the journal significantly influenced the likelihood of race/ethnicity mentions. Case reports in general medicine were the most likely to include race/ethnicity (AOR: 2.18; 95% CI: 2.08 – 2.29), followed by laboratory medicine (AOR: 2.10; 95% CI: 1.94 – 2.28), dentistry (AOR: 1.82; 95% CI: 1.55 – 2.13), and psychiatry (AOR: 1.82; 95% CI: 1.56 – 2.13). A moderate tendency to mention race/ethnicity was also observed in other journal specialties (AOR: 1.37 – 1.97) (Table S1). Surgical specialties were generally less likely to mention race/ethnicity. These included anesthesiology (AOR: 0.27; 95% CI: 0.20 – 0.37), urology (AOR: 0.48; 95% CI: 0.40 – 0.56), traumatology (AOR: 0.59; 95% CI: 0.50 – 0.70), and general surgery (AOR: 0.61; 95% CI: 0.57 – 0.65). Rehabilitation (AOR: 0.31; 95% CI: 0.18 – 0.54) and radiology (AOR: 0.40; 95% CI: 0.35 – 0.47) displayed a strong tendency against reporting race/ethnicity in their clinical cases. Some journal specialties, namely cardiology (AOR: 0.63; 95% CI: 0.56 – 0.72), pneumology (AOR: 0.75; 95% CI: 0.61 – 0.92), and neurology (AOR: 0.79; 95% CI: 0.72 – 0.87), were slightly less inclined to include this information in their clinical case reports.

Finally, the primary diagnosis of the clinical case was also correlated with the likelihood of race/ethnicity mentions, although less strongly than the other variables. Hematological, eye, stomatognathic, metabolic, skin diseases, and infections were significantly associated with slightly higher mentions of race/ethnicity (AOR: 1.20 – 1.32). Conversely, occupational diseases, wounds and injuries, cardiovascular diseases, nervous system diseases, respiratory diseases, and digestive diseases were associated with fewer race/ethnicity mentions (AOR: 0.64 – 0.91).

Detailed information about the AOR of each factor associated with race/ethnicity mentions can be found in Figure 2 and Table S1.

3.4.2. Marital status

Mentions of marital status were notably correlated with several journal specialties such as psychiatry (AOR: 2.6; 95% CI: 1.97 – 3.51), gynecology (AOR: 2.45; 95% CI: 2.01 – 2.99), rehabilitation (AOR: 2.39; 95% CI: 1.31 – 4.35), and forensic medicine (AOR: 2.04; 95% CI: 1.32 – 3.17). Conversely, nephrology (AOR: 0.46; 95% CI: 0.25 – 0.79) and traumatology (AOR: 0.46; 95% CI: 0.26 – 0.79) displayed a pronounced negative correlation with mentions of marital status. Clinical cases pertaining to mental disorders (AOR: 2.14; 95% CI: 1.72 – 2.66) and urogenital diseases (AOR: 1.68; 95% CI: 1.47 – 1.91) were robustly associated with mentions of marital status. Authors from sub-Saharan Africa also exhibited a marked inclination to mention marital status (AOR: 1.98; 95% CI: 1.32 – 2.96).

Several other factors had associations with the likelihood of mentioning marital status, although more moderately. Clinical cases covering a broad spectrum of conditions, such as wounds, neoplasms, infections, digestive, hematological, skin, respiratory, metabolic, musculoskeletal, and nervous diseases, as well as those related to unspecific signs and symptoms, were linked with slightly fewer mentions of marital status (AOR: 0.51 – 0.77). Journals focusing on gastroenterology and general surgery (AOR: 0.53 – 0.74) also demonstrated a subtle association with reduced mentions of marital status.

Lastly, case reports published in the Indian subcontinent or authored by individuals from the Middle East, the Indian subcontinent, North Africa, and Southeast Asia were more inclined to mention marital status (AOR: 1.31 – 1.75). Further details on marital status mentions can be found in Figure 3 and Table S2.

3.4.3. Sexual orientation

The mention of sexual orientation was profoundly correlated with the diagnosis of infectious diseases (AOR: 25.00; 95% CI: 19.68 – 31.75). Other robustly associated factors include case reports published in South America (AOR: 4.04; 95% CI: 1.92 – 8.50) and North America (AOR: 2.15; 95% CI: 1.31 – 3.55). In contrast, journal specialties, such as pediatrics (AOR: 0.16; 95% CI: 0.07 – 0.39) and surgery (AOR: 0.46; 95% CI: 0.30 – 0.69), demonstrated a strong negative correlation with mentions of sexual orientation. A similar trend was also observed across a variety of diagnoses, including cardiovascular, musculoskeletal, and respiratory (AOR: 0.26 – 0.37).

Authors from the Indian subcontinent (AOR: 0.16; 95% CI: 0.05 – 0.51) and the Middle East (AOR: 0.16; 95% CI: 0.04 – 0.63) were considerably less inclined to mention sexual orientation. Conversely, authors from North America (AOR: 1.47; 95% CI: 1.13 – 1.91) and Western Europe (AOR: 1.46; 95% CI: 1.15 – 1.87) were more inclined to mention sexual orientation more frequently than the authors from other regions. Further details on

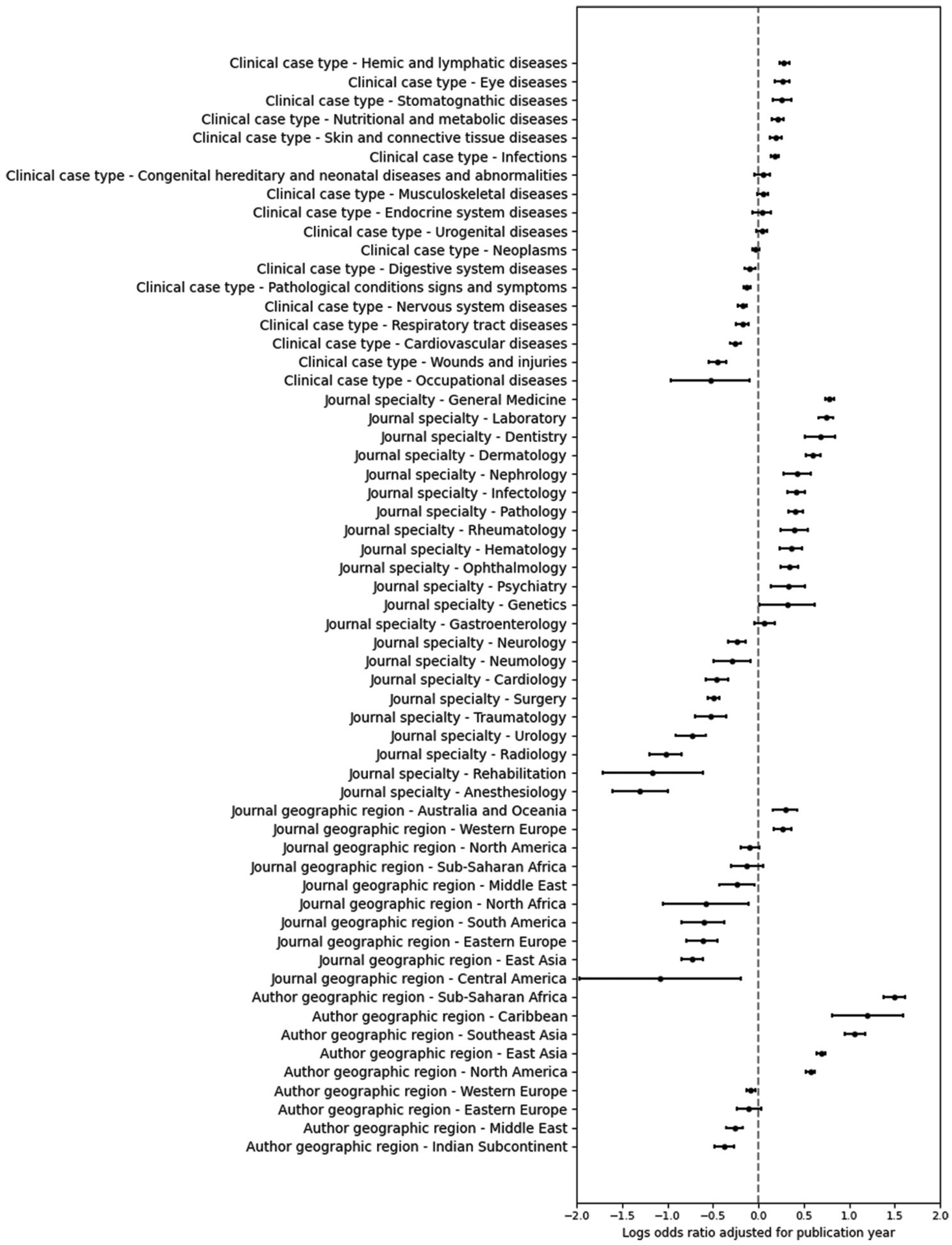


Figure 2. Adjusted odds ratios for the probability of mentioning race/ethnicity based on clinical case type, journal specialty, journal's geographic region, and author's geographic region. The figure was plotted with Matplotlib.

sexual orientation mentions can be found in Figure 4 and Table S3.

3.4.4. Immigrant status

Mentions of immigrant status were strongly associated with infectious diseases (AOR: 4.51; 95% CI: 3.53 – 5.77) and to a lesser extent, with mental disorders (AOR: 2.05; 95% CI: 1.14 – 3.71). Mentions of immigrant status were also positively and significantly associated with journals specializing in gynecology (AOR: 4.25; 95% CI: 2.64 – 6.82) and psychiatry (AOR: 3.94; 95% CI: 1.95 – 7.95), case reports published in the Middle East (AOR: 2.20; 95% CI: 1.19 – 4.07), and authors from Australia and Oceania (AOR: 2.17; 95% CI: 1.14 – 4.12).

Conversely, reduced mentions of immigrant status were associated with authors from the Indian subcontinent (AOR: 0.09; 95% CI: 0.01 – 0.62) and East Asia (AOR: 0.23; 95% CI: 0.12 – 0.45), case reports published in East Asia

Figure 3. Adjusted odds ratios for the probability of mentioning marital status based on clinical case type, journal specialty, journal's geographic region, and author's geographic region. The figure was plotted with Matplotlib.

(AOR: 0.23; 95% CI: 0.15 – 0.65), and journals specializing in ophthalmology (AOR: 0.12; 95% CI: 0.02 – 0.92) and dermatology (AOR: 0.28; 95% CI: 0.09 – 0.90). Diagnoses pertaining to cardiovascular diseases (AOR: 0.43; 95% CI: 0.27 – 0.69) and neoplasms (AOR: 0.43; 95% CI: 0.31 – 0.63) also displayed marked negative associations with immigrant status mentions. Both general medicine journals (AOR: 1.75; 95% CI: 1.34 – 2.29) and authors from North America (AOR: 1.53; 95% CI: 1.17 – 2.01) demonstrated moderate positive associations with mentions of immigrant status. Further details on immigrant status mentions are available in Figure 5 and Table S4.

3.4.5. Homelessness

Mentions of homelessness were strongly associated with journals in the field of forensic medicine (AOR: 14.92; 95% CI: 5.48 – 40.64). Other strongly correlated factors included journals in the areas of pathology (AOR: 3.95; 95% CI: 1.39 – 11.28) and infectious diseases (AOR: 3.75; 95% CI: 1.77 – 7.94), publications from Eastern Europe (AOR: 4.76; 95% CI: 1.88 – 12.03), and diagnoses related to infections (AOR: 6.36; 95% CI: 3.57 – 11.32), mental disorders (AOR: 5.80; 95% CI: 2.26 – 14.89), and injuries (AOR: 4.73; 95% CI: 2.29 – 9.77). Further information on homelessness mentions is available in Figure 6 and Table S5.

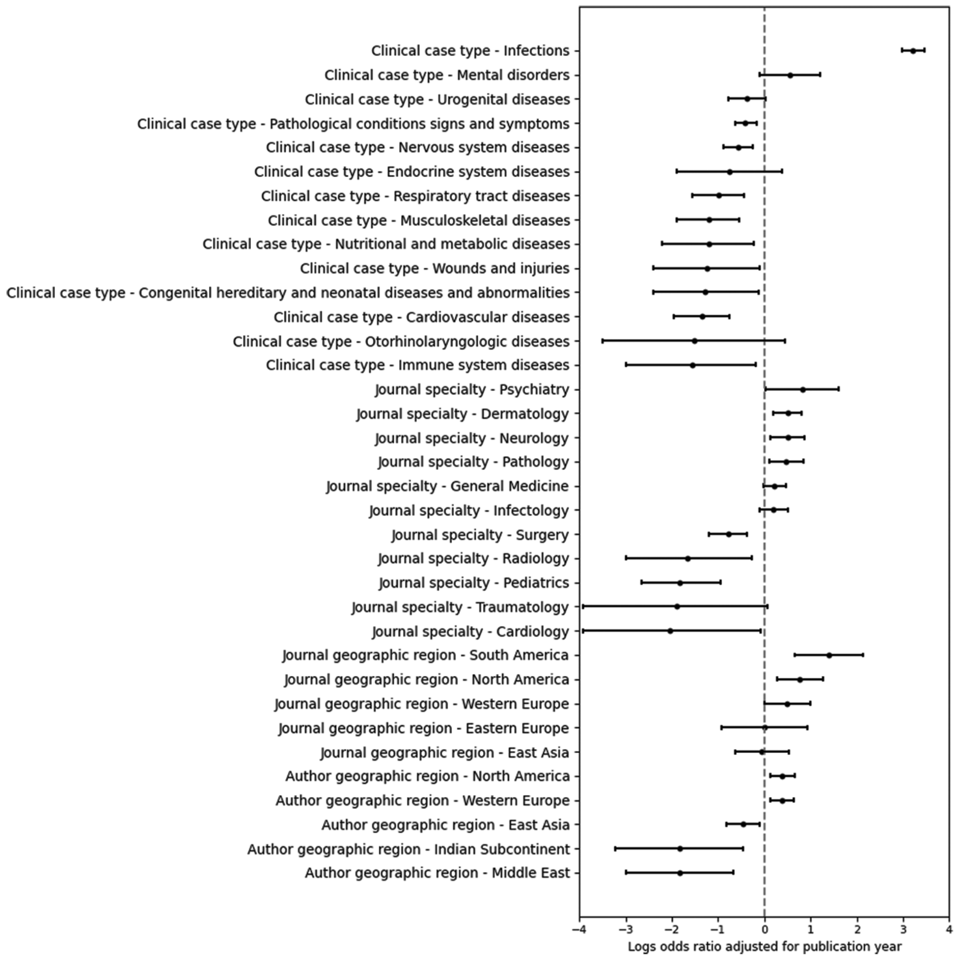


Figure 4. Adjusted odds ratios for the probability of mentioning sexual orientation based on clinical case type, journal specialty, journal's geographic region, and author's geographic region. The figure was plotted with Matplotlib.

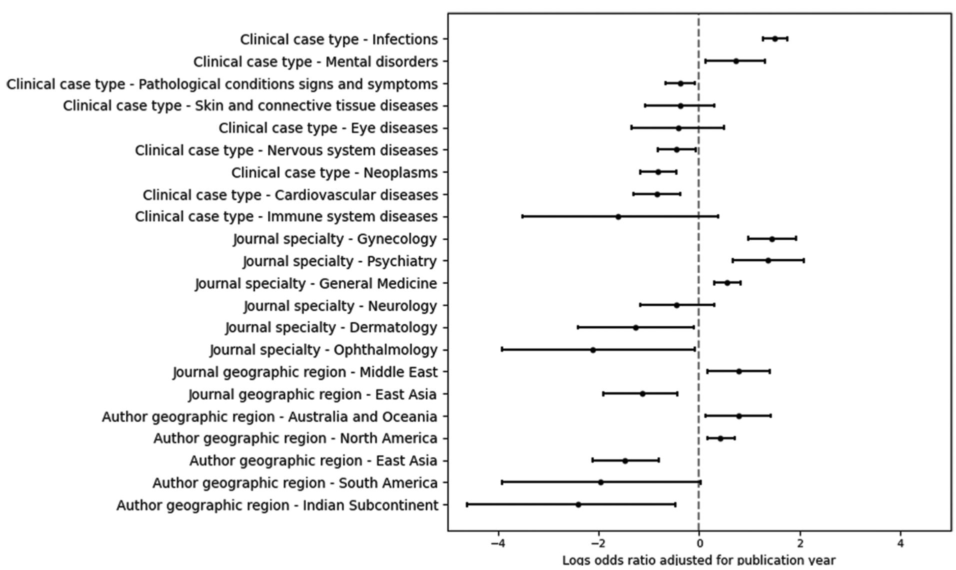


Figure 5. Adjusted odds ratios for the probability of mentioning immigrant status/population group based on clinical case type, journal specialty, journal's geographic region, and author's geographic region. The figure was plotted with Matplotlib.

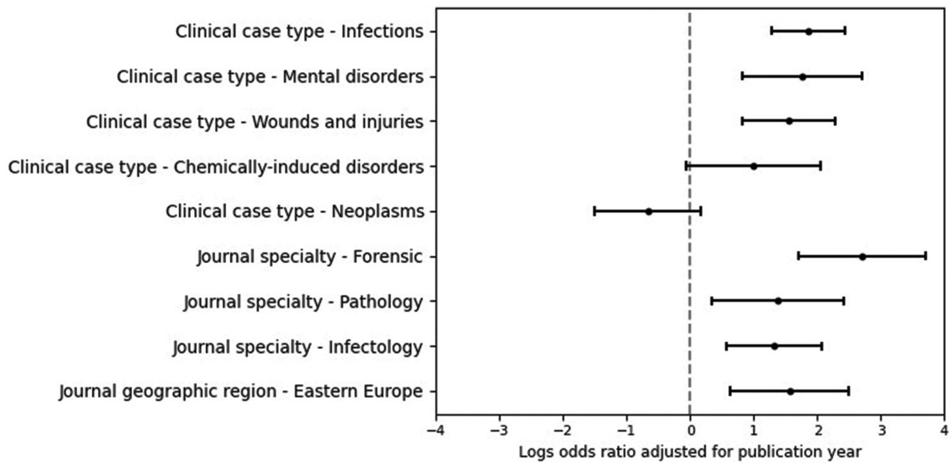


Figure 6. Adjusted odds ratios for the probability of mentioning homelessness/housing based on clinical case type, journal specialty, and journal's geographic region. The figure was plotted with Matplotlib.

3.4.6. Spiritual beliefs

Mentions of spiritual beliefs were strongly correlated with authors from sub-Saharan Africa (AOR: 9.17; 95% CI: 2.84 – 29.64) and the Indian subcontinent (AOR: 4.09; 95% CI: 1.83 – 9.15), journals in the field of psychiatry (AOR: 7.61; 95% CI: 2.93–19.79), publications from the Middle East (AOR: 5.05; 95% CI: 1.99 – 12.85), and clinical cases related to endocrine system diseases (AOR: 3.47; 95% CI: 1.38 – 8.68) and mental disorders (AOR: 3.05; 95% CI: 1.27 – 7.31). In contrast, journals in the field of surgery (AOR: 0.23; 95% CI: 0.06 – 0.96) and clinical cases related to neoplasms (AOR: 0.20; 95% CI: 0.08 – 0.50) were associated with lower probabilities of mentioning patients' spiritual beliefs. Further information on spiritual belief is included in Figure 7 and Table S6.

4. Discussion

4.1. Low prevalence of social determinants of health mentions

Our analysis revealed an uneven distribution of SDoH factors, such that three SDoH factors did not display a clear time-dependent trend. Regarding sexual orientation (Figure S3), a brief increase in mentions occurred in the 1980s, peaking at 40/10,000 case reports. However, the mentions of sexual orientation sharply decreased in the 2000s, leveling at 5/10,000 case reports. We theorized that this surge was associated with the AIDS/human immunodeficiency virus (HIV) outbreak in that period.

There was little variation in race/ethnicity mentions with time (until 2011), depicting steadiness at approximately 300/10,000 case reports (Figure S4). However, between 2011 and 2013, race/ethnicity mentions surged to nearly 550/10,000 case reports. This rate has persisted until 2022, indicating a lasting change in awareness or reporting about race/ethnicity. Nonetheless, further studies are warranted to investigate the reason for the observed trend.

Homelessness mentions displayed a slight increase, but the rate was only 1.29/10,000 case reports, contrasting with the estimated US 1-year homelessness prevalence – about 100 times higher.[36]

Collectively, the data revealed no consistent longitudinal SDoH reporting trends. Observable shifts were sporadic, brief, or tied to specific periods, highlighting the variability of SDoH in the medical literature.

4.2. Risk of biases in the social determinants of health

Our findings reported that diagnosis significantly affects SDoH mentions. Both individual cultural norms (reflected by the author's origins) and institutional policies (indicated by the journal's origins and specialties) impacted SDoH mention frequency. Notably, individual regional contexts exhibited distinct patterns when contrasted with institutional regional contexts represented by journals. In addition, a journal's specialty influences SDoH mentions. Specifically, journals on psychiatry, general medicine, and medical specialties tend to mention SDoH more than surgical specialty journals. These findings emphasized the need for a standardized approach to SDoH reporting across varied geographies and specialties.

Notably, our data revealed potential biases in SDoH reporting in the medical literature. Certain SDoH reports, such as sexual orientation with infectious diseases or homelessness with mental disorders, are overemphasized, potentially reinforcing stereotypes or

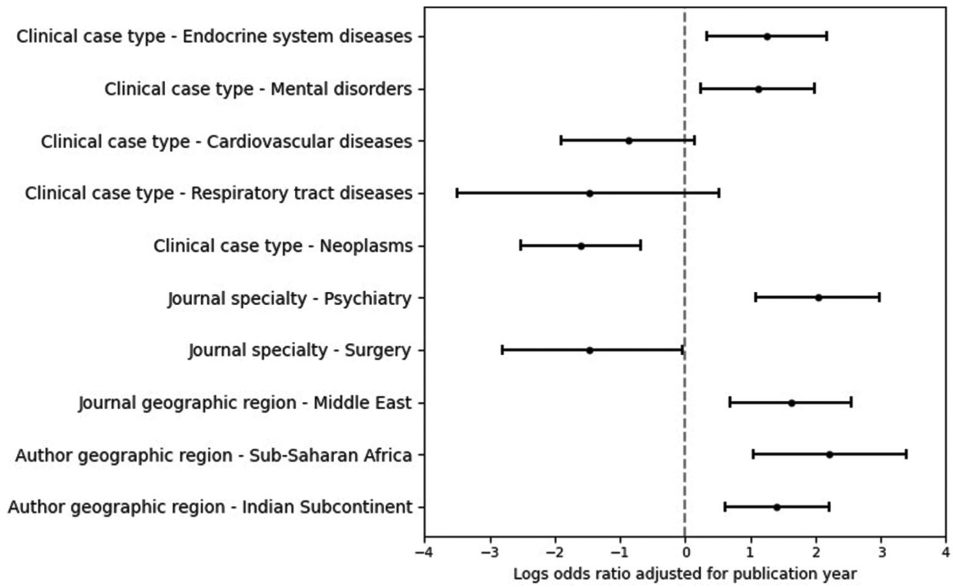


Figure 7. Adjusted odds ratios for the probability of mentioning spiritual beliefs based on clinical case type, journal specialty, journal's geographic region and author's geographic region. The figure was plotted with Matplotlib.

creating oversimplified narratives. Furthermore, these biases risk being duplicated in training large language models, especially those using self-supervised methods with medical literature as data.

4.3. Technological opportunities

Despite the low prevalence of SDoH mentions in clinical case reports, using NER models through Spark NLP offer a potential path for broad-scale clinical record analysis on SDoH mentions. Notably, this method can be used on standard computing hardware,[25] providing access to advanced data analytics. Our research indicated that NER models are more efficient than larger models (e.g., GPT), especially for specific tasks like clinical entity detection. This technology can be used not only for reviewing clinical case reports but also for analyzing EHRs in the search of SDoH,[37,38] thereby enhancing research scalability. In addition, high-level computational analysis could be performed with regular laptops and central processing units (CPUs). Recent studies successfully designed NER models to extract SDoH from clinical narratives.[27] However, the primary objective of our research was not merely to validate these NER models but to analyze the factors associated with the likelihood of mentioning specific SDoH when describing a clinical case.

4.4. Limitations

Our investigation had several limitations that warrant consideration. First, our dataset only included published clinical case reports, which might not reflect the full spectrum of clinical situations or health-care settings. This could lead to a skewed representation of certain regions, affecting our understanding of cultural influences on SDoH mentions.

Second, our analysis might understate SDoH mentions due to two main reasons: our focus was limited to abstracts, specifically sentences outlining primary patient characteristics; and the NER model used had a potential for false negatives, evidenced by the recalls not being 100%. Given the low SDoH mentions in the PubMed corpus, fully evaluating the NER model's recall was challenging. However, our external validation revealed satisfactory recall metrics, and we inferred that the false negatives were likely evenly spread across the model's attribute, subsequently preventing significant impacts on the results from our logistic regression analysis.

In our analysis, we observed that most of the odds ratios (ORs) for the SDoH factors were negative. This finding suggested that specific SDoH mentions within the literature were rare and, when present, were often linked to particular characteristics such as diagnoses, specialties, and cultures. Consequently, this led to OR < 1 for most of the analyzed features. The substantial sample size of our study further amplified the ability of the model to detect statistically significant effects, even for minor associations, adhering to the stringent p-value threshold of $P < 0.0001$.

The prevalence of negative ORs could also be due to overadjustment. Overadjustment occurs when a model includes too many variables or inappropriate variables, leading to biased estimates of the effect size. Despite this risk, the extensive inclusion of variables in our model was a deliberate choice, reflective of the exploratory nature of our research. This project aimed to uncover existing relationships and identify factors potentially associated with SDoH mentions in the literature. To mitigate the risk of arbitrary variable selection, we employed a stepwise approach, including only variables with p-values of 0.001 or less, ensuring that each variable included in the model contributed significantly to the explanatory power of our analysis.

However, we acknowledge that understanding the causality behind these associations requires more sophisticated modeling techniques. Our findings provide the foundation for future research endeavors and in-depth studies that can employ more advanced statistical models to unravel the causal pathways linking SDoH to health outcomes. These studies will be crucial for developing targeted interventions and policies aimed at addressing SDoH more effectively within health-care practices and research.

5. Conclusion

The limited mentions of SDoH in clinical case reports underscore the necessity for better SDoH integration into medical documentation. To mitigate biases in statistical analyses using clinical notes or medical journal content, consistent recording and reporting of SDoH are essential. Spark NLP offers promising avenues for enhancing the extraction and analysis of SDoH from EHRs, highlighting the importance of AI model development to prevent biases that could negatively affect health-care fairness and delivery.

For future research, conducting a similar analysis on the factors associated with SDoH mentions in the full texts of clinical case reports could yield deeper insights. In addition, analyzing actual EHR notes to compare the prevalence and representation of SDoH across different specialties or health-care centers could provide valuable information. Such comparative studies could elucidate the representation and documentation of SDoH across various health-care settings, potentially guiding targeted interventions and policy changes to promote equitable health-care outcomes.

In conclusion, enhancing the documentation and representation of SDoH in the medical literature is critical for advancing toward more informed, equitable, and effective health-care practices and policies. Future studies focused on expanding the scope of analysis to full texts and EHRs could significantly contribute to our understanding and implementation of SDoH in clinical care.

Acknowledgments

None.

Funding

This work has been funded by John Snow Labs Inc.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Julio Bonis, Veysel Kocaman
Formal Analysis: Julio Bonis
Investigation: Julio Bonis, Veysel Kocaman
Methodology: Julio Bonis, David Talby
Writing – Original Draft: Julio Bonis
Writing – Review & Editing: Veysel Kocaman, David Talby

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The dataset utilized for the logistic regression analysis will be made available upon publication. Interested parties can obtain access for academic purposes by directly contacting the authors (julio@johnsnowlabs.com or veysel@johnsnowlabs.com) and signing a data access agreement.

References

1. McGinnis JM, Williams-Russo P, Knickman JR. The case for more active policy attention to health promotion. *Health Aff (Millwood)*. 2002;21:78-93.

   doi: 10.1377/hlthaff.21.2.78

2. Galea S, Tracy M, Hoggatt KJ, DiMaggio C, Karpati A. Estimated deaths attributable to social factors in the United States. *Am J Public Health*. 2011;101:1456-1465.

   doi: 10.2105/AJPH.2010.300086

3. Hatef E, Kharrazi H, Nelson K, *et al*. The association between neighborhood socioeconomic and housing characteristics with hospitalization: Results of a national study of Veterans. *J Am Board Fam Med*. 2019;32:890-903.

   doi: 10.3122/jabfm.2019.06.190138

4. Hood CM, Gennuso KP, Swain GR, Catlin BB. County health rankings: Relationships between determinant factors and health outcomes. *Am J Prev Med*. 2016;50:129-135.

   doi: 10.1016/j.amepre.2015.08.024

5. Walker RJ, Strom Williams J, Egede LE. Influence of race, ethnicity and social determinants of health on diabetes outcomes. *Am J Med Sci.* 2016;351:366-373.

doi: 10.1016/j.amjms.2016.01.008

6. Teshale AB, Htun HL, Owen A, *et al.* The role of social determinants of health in cardiovascular diseases: An umbrella review. *J Am Heart Assoc.* 2023;12:e029765.

doi: 10.1161/JAHA.123.029765

7. Enard KR, Coleman AM, Aver Yakubu R, Butcher BC, Tao D, Hauptman PJ. Influence of social determinants of health on heart failure outcomes: A systematic review. *J Am Heart Assoc.* 2023;12:e026590.

doi: 10.1161/JAHA.122.026590

8. Ludwig J, Sanbonmatsu L, Gennetian L, *et al.* Neighborhoods, obesity, and diabetes - a randomized social experiment. *N Engl J Med.* 2011;365:1509-1519.

doi: 10.1056/NEJMsa1103216

9. Wang M, Pantell MS, Gottlieb LM, Adler-Milstein J. Documentation and review of social determinants of health data in the EHR: Measures and associated insights. *J Am Med Inform Assoc.* 2021;28:2608-2616.

doi: 10.1093/jamia/ocab194

10. Daniel H, Bornstein SS, Kane GC, *et al*. Addressing social determinants to improve patient care and promote health equity: An American college of physicians position paper. *Ann Intern Med.* 2018;168:577-578.

doi: 10.7326/M17-2441

11. Handerer F, Kinderman P, Tai S. The need for improved coding to document the social determinants of health. *Lancet Psychiatry*. 2021;8:653.

doi: 10.1016/S2215-0366(21)00208-X

12. Cottrell EK, Dambrun K, Cowburn S, *et al.* Variation in electronic health record documentation of social determinants of health across a national network of community health centers. *Am J Prev Med.* 2019;57:S65-S73.

doi: 10.1016/j.amepre.2019.07.014

13. Guo Y, Chen Z, Xu K, *et al.* International Classification of Diseases, Tenth Revision, clinical modification social determinants of health codes are poorly used in electronic health records. *Medicine (Baltimore)*. 2020;99:e23818.

doi: 10.1097/MD.0000000000023818

14. Truong HP, Luke AA, Hammond G, Wadhera RK, Reidhead M, Joynt Maddox KE. Utilization of social determinants of health ICD-10 Z-codes among hospitalized patients in the United States, 2016-2017. *Med Care.* 2020;58:1037-1043.

doi: 10.1097/MLR.0000000000001418

15. Hatef E, Rouhizadeh M, Tia I, *et al.* Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: A retrospective analysis of a multilevel health care system. *JMIR Med Inform.* 2019;7:e13802.

doi: 10.2196/13802

16. Guevara M, Chen S, Thomas S, *et al*. Large language models to identify social determinants of health in electronic health records. *NPJ Digit Med*. 2024;7:6.

doi: 10.1038/s41746-023-00970-0

17. Jiménez Carrillo M, Fernández Rodker J, Sastre Paz M, Alberquilla Menendez-Asenjo Á. Does the electronic health record reflect the social determinants of health from primary health care? *Aten Primaria*. 2021;53:36-42.

doi: 10.1016/j.aprim.2020.01.007

18. Gold R, Bunce R, Cowburn S, *et al.* Adoption of social determinants of health EHR tools by community health centers. *Ann Fam Med.* 2018;16:399-407.

doi: 10.1370/afm.2275

19. Tamang S, Humbert-Droz M, Gianfrancesco M, Izadi Z, Schmajuk G, Yazdany J. Practical considerations for developing clinical natural language processing systems for population health management and measurement. *JMIR Med Inform*. 2023;11:e37805.

doi: 10.2196/37805

20. Elbattah M, Arnaud É, Gignon M, Dequen G. The Role of Text Analytics in Healthcare: A Review of Recent Developments and Applications: In: *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies* SCITEPRESS - Science and Technology Publications, Vienna, Austria; 2021. p. 825-832.

doi: 10.5220/0010414508250832

21. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North*. Vol. 1; 2019. p. 4171-4186.

doi: 10.18653/v1/n19-1423

22. Lee J, Yoon W, Kim S, *et al.* BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36:1234-1240.

doi: 10.1093/bioinformatics/btz682

23. Haq HU, Kocaman V, Talby D. Mining adverse drug reactions from unstructured mediums at scale. In: *Shaban-Nejad A, Michalowski M, Bianco S, editors. Multimodal AI in Healthcare. Studies in Computational Intelligence*. Vol. 1060. Cham: Springer; 2023.

doi: 10.1007/978-3-031-14771-5_26

24. Kocaman V, Talby D. Accurate clinical and biomedical named entity recognition at scale. *Softw Impacts*. 2022;13:100373.

doi: 10.1016/j.simpa.2022.100373

25. Kocaman V, Talby D. Spark NLP: Natural language understanding at scale. *Softw Impacts*. 2021;8:100058.

doi: 10.1016/j.simpa.2021.100058

26. Zhu Y, Yuan H, Wang S, *et al.* Large language models for information retrieval: A survey; 2023.

doi: 10.48550/ARXIV.2308.07107

27. Raza S, Dolatabadi E, Ondrusek N, Rosella L, Schwartz B. Discovering social determinants of health from case reports using natural language processing: Algorithmic development and validation. *BMC Digit Health*. 2023;1:35.

doi: 10.1186/s44247-023-00035-y

28. Hazlehurst B, Naleway A, Mullooly J. Detecting possible vaccine adverse events in clinical notes of the electronic medical record. *Vaccine*. 2009;27:2077-2083.

doi: 10.1016/j.vaccine.2009.01.105

29. Banerji A, Lai KH, Li Y, *et al.* Natural language processing combined with ICD-9-CM codes as a novel method to study the epidemiology of allergic drug reactions. *J Allergy Clin Immunol Pract.* 2020;8:1032-1038.e1.

doi: 10.1016/j.jaip.2019.12.007

30. Detect Sentences in Healthcare Texts. *Detect Sentences in Healthcare Texts*. John Snow Labs Inc. Available from: https://nlp.johnsnowlabs.com/2021/08/11/sentence_detector_dl_healthcare_en.html [Last accessed on 2024 Apr 06].

31. Social Determinants of Health. *Social Determinants of Health*. John Snow Labs Inc.; 2023. Available from: https://nlp.johnsnowlabs.com/2023/06/13/ner_sdoh_en.html [Last accessed on 2024 Apr 06].

32. Kocaman V, Talby D. Biomedical named entity recognition at scale. In: Del Bimbo A, Farinella GM, Escalante HJ, *et al*, editors. *Pattern Recognition. ICPR International Workshops and Challenges*. Vol. 12661. Cham: Springer International Publishing; 2021. p. 635-646.

doi: 10.1007/978-3-030-68763-2_48

33. Brown T, Mann B, Ryder N, *et al*. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*. Curran Associates Inc.: United States; 2020: 1877-1901.

doi: 10.48550/ARXIV.2005.14165

34. *OpenAI. GPT-4 Technical Report*; 2023.

doi: 10.48550/ARXIV.2303.08774

35. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. *Biom J.* 2018;60:431-449.

doi: 10.1002/bimj.201700067

36. Tsai J. Lifetime and 1-year prevalence of homelessness in the US population: Results from the national epidemiologic survey on alcohol and related conditions-III. *J Public Health Oxf Engl.* 2018;40:65-74.

doi: 10.1093/pubmed/fdx034

37. Lybarger K, Dobbins NJ, Long R, *et al.* Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *J Am Med Inform Assoc.* 2023;30:1389-1397.

doi: 10.1093/jamia/ocad073

38. Stewart De Ramirez S, Shallat J, McClure K, Foulger R, Barenblat L. Screening for social determinants of health: Active and passive information retrieval methods. *Popul Health Manag.* 2022;25:781-788.

doi: 10.1089/pop.2022.0228

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Enhancing patient safety through integrated sensor technology and machine learning for bed-based patient movement detection in inpatient care

Jonathan Mayer[1], Rejath Jose[1], Molly Bekbolatova[1], Chris Coletti[2], Timothy Devine[2], and Milan Toma[1]*

[1]Department of Osteopathic Manipulative Medicine, New York Institute of Technology College of Osteopathic Medicine, Old Westbury, New York, United States of America

[2]Ferrara Center for Patient Safety and Clinical Simulation, New York Institute of Technology College of Osteopathic Medicine, Old Westbury, New York, United States of America

***Corresponding author:**
Milan Toma
(tomamil@tomamil.com)

Citation: Mayer J, Jose R, Bekbolatova M, Coletti C, Devine T, Toma M. Enhancing patient safety through integrated sensor technology and machine learning for bed-based patient movement detection in inpatient care. *Artif Intell Health.* 2024;1(2): 132-143. doi: 10.36922/aih.2790

Received: January 19, 2024

Accepted: March 27, 2024

Published Online: April 23, 2024

Abstract

The occurrence of inpatient falls and new-onset seizures are common complications during hospital stays, posing risks to patient safety and potentially leading to prolonged hospital stays and further complications. Given the constraints on medical staff's ability to provide constant monitoring due to their workload, the implementation of a sensor device equipped with machine learning capabilities to recognize and prevent these events becomes imperative. This study utilized data acquired through the Movella Xsens sensor, which detects real-time motions and 3D movements, in conjunction with the PyCaret machine-learning algorithm. Adult-sized and infant-sized mannequins were used to assess the algorithm's ability in predicting specific movements associated with breathing, seizures, rolling to the right side, rolling to the left side, rolling off the bed from the left, and rolling off the bed from the right. The study achieved an overall 89% accuracy rate in detecting each specific movement using the combination of PyCaret and Xsens sensors. The application of PyCaret alongside Xsens sensors demonstrates promising results in accurately detecting movements, thereby mitigating falls and post-seizure complications in an inpatient setting, consequently improving patient safety. Further exploration of this technology holds the potential to revolutionize healthcare delivery by incorporating it into a trigger alert system capable of promptly warning medical staff of urgent situations through real-time capture and analysis of potentially harmful motions.

Keywords: Inpatient falls; Sensor device; Machine learning; Patient safety; Movement detection

1. Introduction

During inpatient hospital stays, falls serve as prominent starting points for numerous significant afflictions in patients. Annually, the documentation records up to 1 million inpatient hospital falls, with nearly 250,000 causing various injuries and 11,000 even leading to fatalities.[1] Inpatient hospital falls not only concern patients on an individual level but also pose detrimental challenges to hospital administrations and insurance companies, as they not only lead to injuries and increased risk of fatal events but simultaneously extend hospital stays and inflate medical care costs. According to the Center for Disease Control's National Center for Injury Prevention and Control, unintended injuries are responsible for more years of potential life lost than any other cause of death; among the reported 3.4 million unintended injuries, 72,000 are attributed to falls.[2] While inpatient hospital falls are objectively viewed as preventable events thus far, there remains a dearth of effective preventive measures.[3,4] Current methods include providing patients with educational videos on fall prevention, deploying various forms of bed alarms, and employing video monitoring in patient rooms.[5,6] Falls occur for a variety of reasons. Accidental falls occur when patients slip, trip, or encounter other environmental factors. Anticipated physiological falls can be best described as falls experienced by patients predisposed to falling, influenced by factors such as previously recorded falls, an inaccurate self-assessment of capabilities, the presence of intravenous lines or saline locks, or the use of an ambulatory aid.[7,8] Anticipated physiological falls constitute the majority of inpatient falls, while unanticipated physiological falls are less frequent and unpredictable.[8] This study aims to address the longstanding challenge of hospital patient falls by presenting a solution.

Hospital-onset seizures, defined as seizures occurring in hospitalized patients not admitted for seizure-related incidents and lacking a history of seizures, represent jarring occurrences often associated with extended hospital stays and heightened medical care requirements.[9] A previous study investigating hospital-onset seizures identified 218 patients, revealing that 33% experienced generalized tonic-clonic seizures, while metabolic derangements accounted for 25% of the remaining cases.[9] In addition, the study discovered a higher incidence of mortality among patients experiencing hospital-onset seizures compared to those with preexisting histories of seizures, with rates of 19% and 5%, respectively.[9] Thus, hospital-onset seizures typically manifest as new-onset and often recur, coinciding with elevated mortality rates. The results gathered in this study on patient seizures propose a potential novel safety measure for early seizure detection and swift intervention.

PyCaret is a low-code, open-source machine learning library within Python designed to streamline coding efforts while increasing the time available for analysis. Its application has extended to evaluating turnaround time, a critical performance indicator in medical diagnostic laboratories.[10] In addition, PyCaret has demonstrated promise in studies focusing on histological variants of bladder and urothelial carcinomas.[11] Notably, PyCaret has aided in predicting the evolution of mild cognitive impairment to Alzheimer's disease.[12] In recent years, machine learning has seen growing use in the healthcare industry with the objective of enhancing patient results and fostering more effective and tailored care practices.[13,14] In this study, PyCaret was used to classify data surrounding both simulated patient falls and simulated patient seizures, specifically classifying data relating to six particular motions: breathing, seizures, rolling to the right side, rolling to the left side, rolling off the bed from the left, and rolling off the bed from the right. This study serves as an innovative approach to fall prevention that has not been previously implemented in hospitals. Using the Movella Xsens motion sensors to continuously gather continuous data on patient movements and employing machine learning algorithms to classify said data, trigger alert systems for hospital staff may be developed. The goal is to prevent adverse hospital events such as falls and hospital-onset seizures, thereby leading to better patient outcomes and improved patient safety.

This study directly tackles the significant challenge posed by inpatient falls and hospital-onset seizures, occurrences that not only jeopardize patient safety but also impose considerable costs on healthcare systems. Despite existing preventative measures, these events remain a concern. In response, this study introduces a novel approach by utilizing the Movella Xsens sensors alongside the PyCaret machine learning algorithm to predict and potentially prevent such incidents. The sensor device detects real-time motions, while the PyCaret algorithm classifies these movements to recognize patterns associated with risk events. This integrated approach was tested using mannequins, demonstrating an 89% accuracy in movement detection. The findings suggest the potential of this technology to serve as an effective alert system, thereby advancing patient safety by enabling timely interventions by medical staff.

2. Materials and methods

This section provides an overview of the experimental data collection process, operationalization of sensors and movements to obtain relevant data, data preprocessing for machine learning analysis, and utilization of the PyCaret machine learning library to establish, analyze, and evaluate classification models. It emphasizes the metrics used to determine the success and reliability of the models in predicting different types of patient movement, ultimately contributing to the study's goal of improving patient safety through early detection of fall or seizure events.

In conducting this study, a methodology was utilized to replicate the movements associated with inpatient falls and seizures as follows: Adult-sized and infant-sized mannequins were employed to represent a range of patient demographics, ensuring that the collected movement data spanned various relevant physiologies for our algorithm's predictive capabilities. The use of such mannequins allows for consistent and repeatable movement simulations, which are crucial for machine learning applications. The rationale behind selecting the PyCaret machine learning library is twofold. First, PyCaret's low-code environment significantly streamlines the development process, thereby facilitating a more efficient exploration of different predictive models. Second, it offers a comprehensive suite of evaluation metrics and algorithms suitable for both binary and multiclass classification problems, making it particularly well-suited for the complex task of classifying the nuanced movements indicative of potential falls or seizures. This adaptability and ease of use render PyCaret highly suitable for health-care settings, where rapid and accurate decision-making is paramount for patient safety.

This section also provides detailed descriptions of the approaches used for data collection, preprocessing, and the setup of the machine learning model as follows:

(i) Data collection: This segment describes the use of Xsens DOT sensors to gather real-time motion data reflecting 3D orientations in space, which is capable of detecting Euler angles in the X-, Y-, and Z-axes. It discusses the placement of the sensor and the mechanics of its securement to the mannequin's chest. The methodology elucidates the specific movements imitated by the mannequins (e.g., breathing, seizures, rolls, and falls) to collect diverse movement data while distinguishing between the use of adult and infant mannequins for different movements.

(ii) Data preprocessing: This section outlines the process of managing raw datasets, which involves the segregation of collected data into subsets correlating to specific movements of interest. Focus is given to the significance of the Euler angle points in the X-axis and the quantification method for capturing distinct movement data, including simulated rolling and falling off a bed by a mannequin.

(iii) Machine learning model setup: Details are provided on the utilization of PyCaret, a supervised machine-learning module for the study, emphasizing its streamlined workflow and five key steps: setup, compare models, analyze model, save model, and prediction. The process of setting up PyCaret, providing data, labeling the target, and ensuring reproducibility through session IDs is described. Detailed information about the dataset, including data shape before and after transformations and division into training and test sets, is provided.

(iv) Evaluation metrics: PyCaret-generated metrics such as accuracy, area under the curve (AUC), recall, precision, F1 score, and others are introduced. The definitions and significance of these metrics for model evaluation are articulated, including the intricacies of how performance metrics, such as accuracy, precision, recall, and F1 score, are calculated. Other important metrics, such as the receiver operator characteristic curve (ROC), AUC, Cohen's Kappa, Matthews correlation coefficient (MCC), and training time (TT), are discussed, explaining each metric's value range and its implications for the model's predictive performance.

2.1. Data collection

The data were collected using the Xsens DOT sensors capable of capturing Euler angles in the X-, Y-, and Z-axes, also known as roll, pitch, and yaw, respectively, to depict real-time 3D orientation in space. In addition, these sensors have demonstrated efficacy in capturing distinctive data related to various patient movements.[15] As illustrated in Figure 1, one sensor was placed on the mannequin's chest, specifically at the center of the sternum, to evaluate the aforementioned six movements of interest. The sensor was securely affixed to the mannequin using duct tape arranged in a cross-shaped configuration. Subsequently, it was wirelessly connected through Bluetooth through the Movella Dot App. The application allows for continuous streaming and data collection once initiated by the user. The sensors were activated and deactivated for each overall movement of interest, with the collected data immediately transferred to the device connected to the sensor. To collect the data on breathing and seizures, an adult mannequin was used. Conversely, an infant mannequin was used to collect data on the remaining four movements. Data pertaining to breathing involved the mannequin performing one full cycle of tidal volume inhalations and exhalations continuously for 3 min. Seizure data were collected by inducing a seizure in the mannequin for 10 min. For the collection of data on rolling to the side,

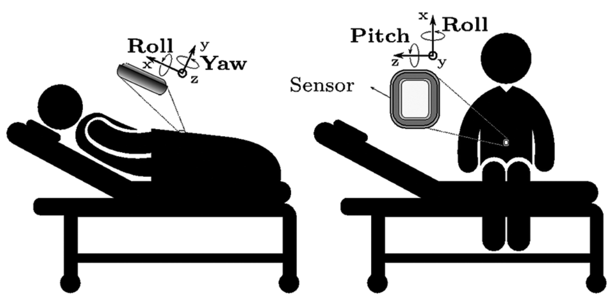


Figure 1. Depiction of the sensor placement and the respective angular motions. Image created using Inkscape

the infant mannequin was initially positioned in the supine position and then rolled approximately 90 degrees to the left side and back to the original supine position, a procedure repeated 100 times. Similarly, the same procedure was performed for rolling to the right side. Data regarding falling off the bed from the left side were collected by initially positioning the infant mannequin in the supine position and subsequently rolling it beyond its left side until it fell off the bed, a procedure repeated 100 times. Given the near-identical nature of dropping the mannequin from its right side, this movement was not performed.

2.2. Data preprocessing

The raw dataset for each respective movement, collected by the sensors, was immediately gathered following the completion of repeated movements. These datasets were divided into six specific movements of interest, each ranging from approximately 18,000 to 35,000 data points containing the Euler angle points in the X-, Y-, and Z-axes. Of particular interest were the Euler angle points along the X-axes exclusively. To isolate the data for each movement, the raw dataset needed to be subdivided to capture data for every 100 movements. Approximately 200 data points were needed to represent one complete movement. For example, the information of one complete roll to the left side consists of 200 points, capturing the transition from the starting supine position to rolling the mannequin to its left side and back to the starting supine position. This collection of 200 points was repeated to ensure a clear delineation of values for each of the 100 movements. This data segmentation process was similarly applied to the other movements of interest. To mimic the mannequin dropping from the right side, the values obtained from dropping the mannequin from the left side were negated. Once the data for each completed movement were collected, it was transferred to one single Excel sheet for further analysis. From the raw datasets, 100 movements were collected for each roll to the right side, roll to the left side, and seizures. Ninety-five movements were collected for each dropping off the bed from the left and right sides, while 89 movements were collected for breathing.

2.3. Pycaret setup

The classification module in PyCaret is a supervised machine learning module designed for classifying elements and aiming to predict categorical class labels that are discrete and unordered. It can handle both binary and multiclass problems, finding applications in various scenarios. The typical workflow in PyCaret for classification consists of five steps: setup, compare models, analyze model, save model, and prediction. The first step, "Setup," initializes the training environment and creates a transformation pipeline. It requires two mandatory parameters, "data" and "target," and offers several optional parameters for customization. The user provides the data in a cohesive fashion with the target labeled appropriately, typically in comma-separated values file (CSV) format. Since this is a classification model, the target is a categorical variable represented numerically (i.e., "Roll right" is 0, "Roll left" is 1, "Drop right" is 2, "Drop left" is 3, "Breathing" is 4, and "Seizure" is 5). The code base for the Google Colaboratory notebook is available at: Sensor_Classification.ipynb.

Once the setup is executed successfully, it displays an information grid with experiment-level details (Table 1). The session ID is a pseudo-random number (123 in this case) used as a seed for reproducibility in all functions throughout the PyCaret pipeline, ensuring consistent results when running the same code with the same session ID. The target refers to the column in the dataset (the CSV file) that will be predicted. In this case, the target is designated "Predict." The target type specifies the nature of the target variable, which in this case is "Multiclass," indicating that the target variable has multiple distinct classes for multiclass classification. The original data shape shows the dimensions of the dataset before any transformations, with 579 rows and 203 columns. Similarly, the transformed data shape also has 579 rows and 203 columns, indicating that the dataset was not modified during the setup process. The transformed training set shape indicates that the training dataset contains 405 rows and 203 columns after preprocessing, which was used to train the machine learning models. The transformed test set shape indicates that the test dataset contains 174 rows and 203 columns after preprocessing, which was used to evaluate the performance of the trained models. Therefore, a split of 70% for training and 30% for testing was used.

Table 1. Experiment setup details

No.	Description	Value
0	Session ID	123
1	Target	Predict
2	Target type	Multiclass
3	Original data shape	(579, 203)
4	Transformed data shape	(579, 203)
5	Transformed train set shape	(405, 203)
6	Transformed test set shape	(174, 203)
7	Numeric features	202
8	Preprocess	True
9	Imputation type	Simple
10	Numeric imputation	Mean
11	Categorical imputation	Mode

Numeric features represent columns with numerical values, encompassing both continuous and discrete data. In this context, the dataset comprises 202 numeric features, with the "Predict" feature serving as a categorical target variable (Table 1). The value "True" for preprocessing indicates that preprocessing steps are applied to the data during the setup process. For the current study, preprocessing steps known as "LabelEncoder" and "SimpleImputer" were applied. The label encoder is a preprocessing step applied to convert categorical target variables (if any) into numerical format. It transforms categorical labels into integer values, making them suitable for training the machine learning model. The simple imputer is also a preprocessing step applied to handle missing values in the dataset. It fills in missing values using simple strategies such as the feature's mean, median, or most frequent value. In cases where numeric features have missing values, the "mean" imputation method is used, replacing the missing numeric values with the mean of the corresponding feature. Conversely, for categorical features with missing values, the "mode" imputation method is used, replacing the missing categorical values with the mode (most frequent category) of the corresponding feature.

The "Compare Models" function trains and evaluates the performance of all available estimators using cross-validation, providing a scoring grid with average cross-validated scores. To analyze the performance of a trained model on the test set, the "plot_model" function can be used. It offers different plot types, such as confusion matrix and AUC, for assessing model performance. In certain cases, re-training the model may be required for plotting specific visualizations. Finally, the model with the entire pipeline is saved on disk for future use, especially for prediction of unseen data.

Hence, the typical workflow in PyCaret for a classification task involves several steps, beginning with the "Setup." During "Setup," the user initiates the training environment by defining the dataset (data) and the variable to be predicted (target). In this case, the target refers to various movements like "Roll right," "Roll left," "Drop right," "Drop left," "Breathing," and "Seizure," encoded numerically from 0 to 5, respectively. The following is how PyCaret handles the classification workflow:

- Session ID: In the setup stage, specifying a session ID as a pseudorandom number (e.g., 123) serves as a seed for all randomness within the pipeline, ensuring that the experiment is reproducible. This setup process implies that the random division of data into folds when applying cross-validation or the random selection of data points if any undersampling or oversampling is performed would yield consistent results each time the code is run with the same session ID.
- Data format and preparation: The input data are provided as a CSV file, which is a standard, easy-to-work-with data format. The data include both the features (e.g., sensor readings) and the target. The features are the inputs that the model will learn from, while the target is the output category that the model is trained to predict.
- Target variable: The target variable is categorical, meaning that it does not have a natural order or numerical value; the assigned numbers are just labels for the classes. As the target is represented numerically, each number corresponds to a discrete category of patient movement, and the model learns to predict these categories.
- Workflow steps: The typical workflow in PyCaret for classification consists of five steps: setup, compare models, analyze model, save model, and prediction.
  - (i) Setup: This crucial first step initializes the analysis environment by setting up the data and defining the target. It also performs basic processing like handling missing values, encoding categorical variables, normalizing the data, and potentially feature engineering.
  - (ii) Compare models: This step systematically trains and evaluates different machine learning models using the preprocessed data, subsequently ranking them according to a chosen evaluation metric, usually accuracy for classification tasks.
  - (iii) Analyze model: For the chosen model, its performance metrics, decision boundary, feature importance, confusion matrix, and other insights are analyzed to understand how well the model works. This step provides information about the classifier's behavior under various conditions through ROC curves, precision-recall curves, and classification errors, allowing the user to deeply interrogate specific models and understand areas for improvement.
  - (iv) Save and predict model: With the model saved, predictions can be made on new data that the model has not seen before. This is the ultimate goal of the machine learning workflow—applying the constructed model to make accurate classifications on real-world data.

The training and test datasets are created during the setup, with PyCaret automatically splitting the input data into these subsets. The typical default splits allocate 70% of the data for training and 30% for testing. The session ID ensures consistency in any randomization during this process across sessions. Therefore, the transformed train set shape and test set shape reported after setup refers to the shapes of these datasets post- and pre-processing and splitting. The workflow encapsulated by the PyCaret setup supports the end-to-end process of building and deploying classification models. In the study context, the models are tasked with classifying patient movements based on sensor data. The workflow enables the use of sophisticated machine learning algorithms without requiring the user to dive deep into the algorithmic complexities associated with each model. Consequently, researchers and practitioners can focus more on interpreting the results and less on managing the workflow mechanics.

2.4. Evaluation metrics

Pycaret provides a range of metrics, including precision, recall, F1 score, accuracy, AUC, Cohen's Kappa, MCC, and TT. Accuracy, as defined in Equation I, represents the proportion of correctly classified values out of the total number. Precision, as per Equation II, is computed as the ratio of true positive instances to all predicted positive instances, where a higher precision score indicates fewer false positive predictions. Recall, defined in Equation III, assesses the ability to identify actual positives and is also known as sensitivity. The F1 score, as per Equation IV, synthesizes precision and recall into a single value between 0 and 1; higher scores indicate better performance in both areas, while lower scores suggest poor precision or recall.

$$\text{Accuracy}=\frac{\text{True Positive}+\text{True Negative}}{\text{True Positive}+\text{False Positive}+\text{True Negative}+\text{False Negative}} \tag{I}$$

$$\text{Precision}=\frac{\text{True Positive}}{\text{True Positive}+\text{False Positive}} \tag{II}$$

$$\text{Recall}=\frac{\text{True Positive}}{\text{True Positive}+\text{False Negative}} \tag{III}$$

$$\text{F1 score}=\frac{2}{1/\text{Precision}+1/\text{Recall}}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{IV}$$

The ROC evaluates the difference in the rates of true positive rate and false positive rate results across different decision thresholds.[16] The AUC serves as an indicator of the model's effectiveness, allowing for comparison of performance across various models.[17-19] An AUC equal to one indicates a perfect model, while an AUC exceeding 0.5 indicates that the model's classification capability outperforms random guessing and possesses predictive value. An AUC of 0.5 signifies that the model's classification capacity is equivalent to random guessing, devoid of predictive value. An AUC lower than 0.5 suggests a classification capacity worse than random guessing; however, if a reverse prediction is conducted, it is superior to random guessing. The collection of all sample points forming a line constitutes an ROC curve.[15]

Cohen's Kappa, often known as "Kappa," is a statistical measure used to assess the agreement between predicted and actual classes while also accounting for the level of agreement beyond what would occur by chance. This metric holds particular significance when dealing with imbalanced datasets, as it considers chance-based agreement. Kappa values range from –1 to 1, where 1 signifies perfect agreement, 0 denotes chance-based agreement, and values below 0 indicate predictions worse than random. Meanwhile, MCC serves as another metric for assessing the quality of binary and multiclass classifications; it takes into account true positives, true negatives, false positives, and false negatives, making it useful in scenarios involving imbalanced datasets. Similar to Cohen's Kappa, the MCC also ranges from –1 to 1: a value of 1 indicates flawless prediction capability, while a value of zero represents predictions at random; anything below zero suggests predictive performance worse than random guessing. Finally, TT refers to the duration taken by a specific machine learning model to train on the dataset, typically measured in seconds. This metric offers valuable insight into the time required to train a particular model.

3. Results

A total of 15 machine-learning classification models were tested using Pycaret (Table 2). These models included Light Gradient Boosting Machine (LIGHTLGBM), Extra Tree Classifier (ET), Extreme Gradient Boosting, Random Forest Classifier, Gradient Boosting Classifier, Decision Tree Classifier, K Neighbors Classifier, Naive Bayes, Linear Discriminant Analysis, Logistic Regression, Support Vector Machine - Linear Kernel, Ridge Classifier, AdaBoost Classifier, Quadratic Discriminant Analysis, and Dummy Classifier (DUMMY). DUMMY makes predictions that ignore the input features, serving as a simple baseline for comparison against more complex classifiers.

The LIGHTLGBM model exhibited the highest accuracy, recall, precision, F1 score, Kappa, and MCC. Specifically, the accuracy, recall, precision, F1 score, Kappa, and MCC of the LIGHTLGBM model were 0.89, 0.89, 0.90, 0.89, 0.87, and 0.87, respectively, with an AUC of 0.98. The performance metrics of the LIGHTLGBM model closely resembled those of the ET model, with the ET model displaying slightly lower accuracy, recall, precision, F1, Kappa, and MCC, but marginally higher AUC on the ROC curve (Figure 2). In addition, the confusion matrix

Table 2. Comparisons of accuracy, AUC, recall, precision, F1 score, Kappa, and MCC for different machine learning classifier models

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC	TT (s)
LIGHTLGBM - Light Gradient Boosting Machine	0.8937*	0.9830	0.8937*	0.9017*	0.8926*	0.8723*	0.8744*	3.0400
ET - Extra Trees Classifier	0.8912	0.9856*	0.8912	0.8976	0.8899	0.8693	0.8710	0.2650
XGBOOST - Extreme Gradient Boosting	0.8765	0.9821	0.8765	0.8859	0.8742	0.8517	0.8544	2.0500
RF - Random Forest Classifier	0.8763	0.9853	0.8763	0.8875	0.8737	0.8514	0.8545	0.7430
GBC - Gradient Boosting Classifier	0.8738	0.9799	0.8738	0.8882	0.8717	0.8485	0.8522	13.181
DT - Decision Tree Classifier	0.8273	0.8969	0.8273	0.8493	0.8271	0.7927	0.7974	0.0750
KNN - K Neighbors Classifier	0.7946	0.9456	0.7946	0.8236	0.7943	0.7534	0.7602	0.0570
NB - Naive Bayes	0.7773	0.9589	0.7773	0.7919	0.7739	0.7331	0.7373	0.0480
LDA - Linear Discriminant Analysis	0.7701	0.9248	0.7701	0.7890	0.7690	0.7239	0.7282	0.1420
LR - Logistic Regression	0.7182	0.8895	0.7182	0.7932	0.7175	0.6613	0.6735	1.4500
SVM - SVM – Linear Kernel	0.6270	0.0000	0.6270	0.6821	0.5873	0.5511	0.5844	0.1190
RIDGE - Ridge Classifier	0.5676	0.0000	0.5676	0.6386	0.5372	0.4803	0.5079	0.0910
ADA - Ada Boost Classifier	0.5136	0.7910	0.5136	0.4010	0.4147	0.4120	0.5165	0.5200
QDA - Quadratic Discriminant Analysis	0.3955	0.6350	0.3955	0.5045	0.3514	0.2706	0.3064	0.1170
DUMMY - Dummy Classifier	0.1729	0.5000	0.1729	0.0299	0.0510	0.0000	0.0000	0.0420

Note: *Highest value.
Abbreviations: AUC: Area under the curve; MCC: Matthews correlation coefficient; TT: Training time.

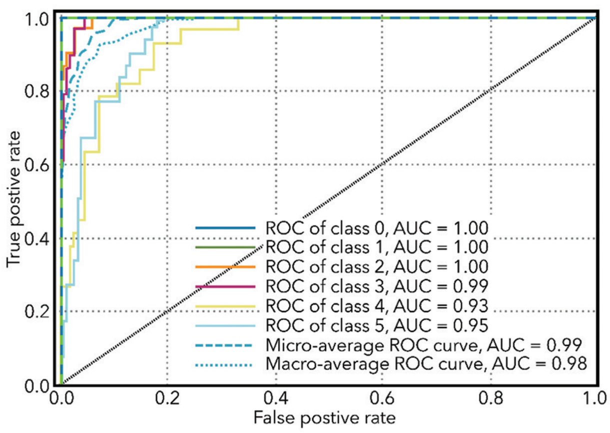


Figure 2. Area under the curves for Light Gradient Boosting Machine (LIGHTLGBM) classifier. Image created with Inkscape
Abbreviations: AUC: Area under the curves; ROC: Receiver operator characteristic curve.

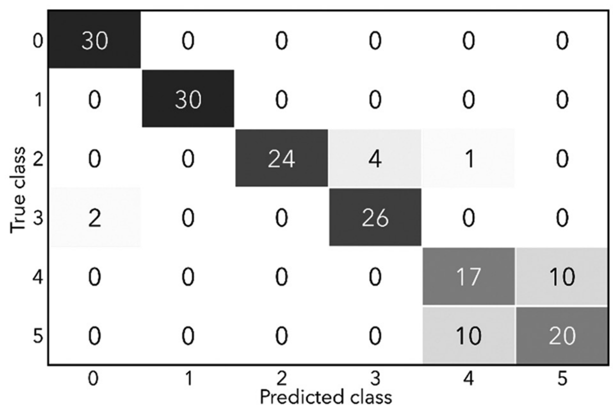


Figure 3. Confusion matrix for the Light Gradient Boosting Machine (LIGHTLGBM) classifier. Image created with Inkscape

generated by Pycaret depicted predictions in the testing split for all categories (Figure 3). A confusion matrix serves as a tool to visualize the performance of a classification model. The diagonal elements of the matrix denote the number of correct predictions for each class, while the off-diagonal elements indicate the number of incorrect predictions, where the model predicts a different class from the actual label. In this study, the confusion matrix is a 6×6 matrix, reflecting the six classes encoded from 0 to 5, for the LIGHTLGBM classifier used.

The numbers on the diagonal can be interpreted as follows: 30 correct predictions for class 0 ("Roll right"); 30 correct predictions for class 1 ("Roll left"); 24 correct predictions for class 2 ("Drop right"); 26 correct predictions for class 3 ("Drop left"); 17 correct predictions for class 4 ("Breathing"); and 20 correct predictions for class 5 ("Seizure"). These numbers indicate that the LIGHTLGBM classifier exhibits the best performance at detecting "Roll right" and "Roll left" movements, as these classes boast the highest number of correct predictions (30 each). The non-zero off-diagonal elements that are 10 or lower represent instances of misclassification of a movement by the model.

For example, a value of 10 in the off-diagonal position indicates that 10 instances of "Breathing" were incorrectly predicted as "Seizure" by the classifier. The implications of these findings from the confusion matrix could be twofold: Classes 0 and 1 are well-recognized by the model, possibly because these movements possess distinct characteristics that are easily discernible by the sensors. Conversely, classes 4 and 5 have the lowest instances of correct prediction (17 and 20, respectively), suggesting that recognizing "Breathing" and "Seizure" movements presents greater difficulty for the model. This challenge could stem from similarities in the patterns of these movements or subtle characteristics that were not captured by the sensors or model features. Misclassifications (off-diagonal numbers) highlight areas where the model confuses one movement for another. Although all off-diagonal numbers are 10 or lower, indicating a decent level of precision, these errors could carry significant consequences depending on the clinical importance of the movements. For instance, mistaking breathing for a seizure (or vice versa) could lead to inappropriate medical interventions or a failure to respond promptly to an actual seizure.

Overall, despite the misclassifications, the numbers suggest a higher number of correct predictions across all classes than incorrect ones, indicating the promising potential of the classifier. However, it is essential to identify the clinical consequences of each type of misclassification to prioritize improvements in the classifier's performance.

4. Discussion

This study, conducted with data collected using Xsens DOT sensors and analyzed utilizing PyCaret, showcases the potential benefits of using artificial intelligence methods in practical, real-life applications. The results indicate an accuracy rate of over 89% in predicting the movements of inpatients based on the provided data. The failure to achieve a perfect accuracy rate from the LIGHTLGBM classifier is attributed to its difficulty in distinguishing between breathing and seizures, with prediction values for these movements approximately 63% and 67%, respectively. This challenge likely stems from the absence of chest movement during seizures, as opposed to the common occurrence of head movements.[20] With that being said, one potential solution to this problem is to place an additional sensor on the mannequin's head to detect seizure movements. The data gathered from this sensor on seizures can be used as the evaluation metrics, while the sensor placed on the chest will solely gather data on the other movements of interest.

While promising, the results of this study do highlight several additional limitations, beyond the one previously mentioned, all of which fortunately are addressable with appropriate tools. One such limitation concerns the utilization of duct tape to secure the monitor to the mannequin's chest. Since duct tape is impractical for live patients, an alternative solution, such as a strap-like device, could be introduced to ensure the monitor remains securely in place on the patient's chest without causing discomfort. Another limitation pertains to the battery life of the sensor device. Throughout the study, it was observed that the sensor's battery depleted to 50% capacity after approximately 3 h of use. This limitation could be resolved by having the patient's nursing team replace the sensor at the 6-h mark. The replaced sensor can then be recharged and reused for subsequent monitoring sessions.

Inpatient falls pose a potential risk of prolonged hospital stays, further injuries, and various complications for many hospitalized patients in recovery. Hospital-onset seizures also present a concern, as they are difficult to predict and can significantly worsen a patient's overall well-being. Early recognition of these events, preferably before their occurrence, is crucial for ensuring patient safety.[21,22] While one-to-one observation is an effective way to monitor high-risk patients, it is not feasible for all admitted patients.[23-26] In addition, while video monitoring appears to be a viable alternative solution, it can raise privacy concerns and incur substantial startup costs and resource investment to install cameras in multiple rooms.[27] The utilization of Euler angle measurements through the Movella Xsens sensors, in conjunction with machine learning algorithms, offers a potential solution to these problems by enabling the detection and classification of specific movements in real time. While there are limitations to using these sensors in practical settings that must still be addressed, this study presents promising results and lays a foundation for further research in this area. By accurately predicting ongoing motion, this novel approach can be incorporated into a trigger alert system for the medical staff, allowing for swift intervention before a fall or seizure occurs. This proactive approach will significantly reduce the risk of adverse events, lower complication rates, and ultimately improve overall patient outcomes.[28-32]

This study presents several limitations. The use of a controlled environment with a limited number of mannequins to replicate patient movements does not fully capture the variability observed in actual patient populations. To better simulate real-world conditions, expanding the study's scope to include a larger and more diverse group of patients across various health-care settings would provide a more comprehensive understanding of sensor capabilities and algorithm performance. In addition, sensors may have limitations in accurately detecting subtle or intricate patient movements, leading to potential false positives or negatives. Therefore, implementing a multimodal sensor approach or utilizing more advanced sensor technologies could enhance the accuracy and reliability of movement detection. Moreover, while machine learning models may perform well under experimental conditions, they might be less effective in real-world applications due to issues such as overfitting or difficulties in interpreting complex data. Thus, employing sophisticated machine learning techniques such as deep learning—which can capture complex patterns—and methods to ensure model robustness against overfitting—such as dropout or data augmentation—is necessary.

In addition, transitioning a model from research settings to widespread clinical use could present challenges due to infrastructure or resource limitations. Developing an adaptable solution necessitates close collaboration with healthcare technology providers while ensuring compatibility across various facility infrastructures. Healthcare providers may hesitate to embrace new technology due to integration challenges with existing workflows or feeling overwhelmed by constant alerts. It is crucial to create technology that seamlessly integrates into current hospital systems and processes. Alert systems need to prioritize important events to minimize unnecessary interruptions and prevent staff fatigue from excessive alerts. Regular maintenance procedures and backup systems must be established to guarantee continuous functionality. Using durable and low-maintenance hardware can also reduce the occurrence of malfunctions. Implementing advanced sensor systems and machine learning models may require an additional investment from healthcare institutions. Therefore, conducting cost-benefit analyses is essential to illustrate the long-term savings associated with reducing patient falls and seizures, such as shorter hospital stays and fewer medical interventions, thereby making a compelling case for investing in technology integration.

The sensor technology collects data on patients' movements, which is then analyzed by a machine learning algorithm to detect falls or other significant events. This system can operate on a closed network within the hospital, without necessarily requiring an external real-time data network connection for its day-to-day functioning. However, it is important to clarify that for the system to be effective in real-world applications; it should enable real-time processing of the sensor data so that immediate alerts can be sent to the medical staff in case of a detected fall or seizure, regardless of whether it is connected to an external network or operates on an internal network. The goal is to ensure timely interventions and improved patient safety, which can be achieved through on-site data processing and alert mechanisms. The actual network requirements would depend on the specifics of the system implementation and operational logistics, such as sensor placement and management of data privacy and security.

In addition, introducing the sensors and machine learning system into a real-world health-care setting would raise considerations regarding data sharing and protection. Implementation would need to comply with data protection regulations, such as the Health Insurance Portability and Accountability Act in the United States or the General Data Protection Regulation in the European Union, which govern the privacy and security of patient data. The hospital's information technology infrastructure would need to ensure adequate measures are in place for data encryption, secure access protocols, and potential anonymization of patient data to prevent unauthorized use or disclosure. Furthermore, the legal aspects of data handling would require careful planning to ensure patient consent is obtained where necessary, and there is transparency in how patient data is used and protected. Failure to adequately address these concerns could potentially jeopardize patient privacy and expose the health-care facility to legal and regulatory risks.

5. Conclusion

The study underscores the potential of combining advanced sensor technology with sophisticated machine learning algorithms to detect and prevent events such as falls and seizures in a hospital setting. Opportunities to enhance the usability and effectiveness of these technologies are evident, particularly in optimizing sensor placement and improving operational logistics, such as battery life management. By addressing these challenges, the approach tested in this study could pave the way for creating robust, real-time monitoring systems that not only alert care providers to potential falls or seizures but also contribute to a broader range of applications in patient care and monitoring. Further exploration and refinement could lead to the development of a more comprehensive solution that mitigates the risks associated with patient falls and seizures, ultimately improving patient outcomes and reducing healthcare costs.

The practical implications of this study, which enhances patient safety with integrated sensor technology and machine learning for bed-based patient movement detection in inpatient care, can significantly affect various aspects of health-care delivery as follows:

(i) Improved patient safety: Accurately detecting movements indicative of potential falls or seizures allows for proactive staff alerts and timely intervention, reducing the incidence and severity of such events and leading to improved patient outcomes.

(ii) Reduced health-care costs: Preventing falls and seizures decreases the average duration of hospital stays and the number of medical interventions, resulting in substantial cost savings for health-care facilities.
(iii) Enhanced monitoring: Continuous and automated monitoring supplements the efforts of health-care staff, allowing them to focus on other critical tasks, knowing that the system will alert them to potential issues with patients.
(iv) Data-driven insights: The data collected by the sensors can provide insights into the most common times or conditions under which falls and seizures occur, facilitating the development of refined care protocols and targeted preventative measures.
(v) Staff efficiency: With a system in place to handle routine monitoring tasks, the staff can allocate their time more efficiently, thereby improving overall productivity.
(vi) Training and education: Data from the sensor technology can serve educational purposes, teaching health-care professionals about patient safety and fall prevention strategies by providing real examples and insights.
(vii) Patient and family peace of mind: Knowing that a sophisticated system monitors their loved ones can provide patients and their families with greater confidence in the care provided by the hospital, potentially improving their overall experience and satisfaction.
(viii) Quality of care metrics: Hospitals can use data from these technologies to demonstrate adherence to patient safety protocols, potentially improving their quality-of-care metrics and accreditation outcomes.
(ix) Improved resource allocation: Predictive analytics enable hospitals to anticipate patient needs and allocate nursing and medical resources more effectively, ensuring that high-risk patients receive more frequent human monitoring compared to low-risk patients.
(x) Legal and compliance benefits: Implementing state-of-the-art technology for patient safety may help health-care facilities comply with legal standards and regulations, potentially reducing the risk of lawsuits associated with inpatient falls and other incidents.
(xi) Scientific advancement: The technology provides a rich data source for further research into patient safety, potentially identifying new risk factors for falls and seizures that were previously unknown.
(xii) Enhanced rehabilitation: For patients recovering from surgery or injury, the technology can monitor rehabilitation progress and ensure that movements are within safe parameters, thereby supporting a more effective recovery process.
(xiii) Broad healthcare applications: Although initially focused on inpatient care, such technology could be extended to other settings such as nursing homes, rehabilitation centers, and even home healthcare, expanding its applications on patient safety beyond the hospital.

Implementing such technology requires a thoughtful approach to address potential challenges such as user acceptance, data security, and ethical considerations. However, the benefits mentioned clearly indicate a profound positive impact on patient care, safety, and hospital operations. In conclusion, the integration of advanced sensor technology and machine-learning algorithms in health-care settings holds immense promise for improving patient safety, thus warranting further research in this technology.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: All authors
Investigation: All authors
Methodology: All authors
Writing – original draft: All authors
Writing – review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The data generated by the sensors that support the findings of this study are available from the corresponding author on reasonable request.

References

1. LeLaurin JH, Shorr RI. Preventing falls in hospitalized patients: State of the science. *Clin Geriatr Med*. 2019;35(2):273-283.

   doi: 10.1016/j.cger.2019.01.007

2. Centers for Disease Control and Prevention. *Web-based*

Injury Statistics Query and Reporting System (WISQARS). National Center for Injury Prevention and Control, Centers for Disease Control and Prevention (Producer); 2023. Available from: https://www.cdc.gov/ncipc/wisqars [Last accessed on 2023 Dec 22].

3. Eost-Telling C, Yang Y, Norman G, *et al*. Digital technologies to prevent falls in people living with dementia or mild cognitive impairment: A rapid systematic overview of systematic reviews. *Age Ageing*. 2024;53:afad238.

   doi: 10.1093/ageing/afad238

4. Wang Y, Jiang M, He M, Du M. Design and implementation of an inpatient fall risk management information system. *JMIR Med Inform*. 2024;12:e46501.

   doi: 10.2196/46501

5. Dykes PC, Burns Z, Adelman J, *et al*. Evaluation of a patient-centered fall-prevention tool kit to reduce falls and injuries: A nonrandomized controlled trial. *JAMA Netw Open*. 2020;3(11):e2025889.

   doi: 10.1001/jamanetworkopen.2020.25889

6. Zubrinic M, Vrbanic L, Keshavjee S. Remote telemonitoring is associated with improved patient safety and decreased workload of nurses. *JTCVS Open*. 2023;16:493-497.

   doi: 10.1016/j.xjon.2023.09.014

7. Morse JM. *Preventing Patient Falls*. Berlin: Springer Publishing Company; 2008.

8. Mcvey L, Alvarado N, Zaman H, *et al*. Interactions that support older inpatients with cognitive impairments to engage with falls prevention in hospitals: An ethnographic study. J Clin Nurs. 2024;33:1884-1895.

   doi: 10.1111/jocn.17006

9. Fields MC, Labovitz DL, French JA. Hospital-onset seizures: An inpatient study. *JAMA Neurol*. 2013;70(3):360-364.

   doi: 10.1001/2013.jamaneurol.337

10. Tsai ER, Demirtas D, Hoogendijk N, Tintu AN, Boucherie RJ. Turnaround time prediction for clinical chemistry samples using machine learning. *Clin Chem Lab Med*. 2022;60(12):1902-1910.

    doi: 10.1515/cclm-2022-0668

11. Evrimler S, Ali Gedik M, Ahmet Serel T, Ertunc O, Alperen Ozturk S, Soyupek S. Bladder urothelial carcinoma: Machine learning-based computed tomography radiomics for prediction of histological variant. *Acad Radiol*. 2022;29(11):1682-1689.

    doi: 10.1016/j.acra.2022.02.007

12. Rye I, Vik A, Kocinski M, Lundervold AS, Lundervold AJ. Predicting conversion to Alzheimer's disease in individuals with Mild Cognitive Impairment using clinically transferable features. *Sci Rep*. 2022;12(1):15566.

    doi: 10.1038/s41598-022-18805-5.

13. Bekbolatova M, Mayer J, Ong CW, Toma M. Transformative potential of AI in healthcare: Definitions, applications, and navigating the ethical landscape and public perspectives. *Healthcare*. 2024;12(2):125.

    doi: 10.3390/healthcare12020125

14. Toma M, Wei OC. Predictive modeling in medicine. *Encyclopedia*. 2023;3(2):590-601.

    doi: 10.3390/encyclopedia3020042

15. Mayer J, Jose R, Kurgansky G, *et al*. Evaluating the feasibility of euler angles for bed-based patient movement monitoring. *Signals*. 2023;4(4):788-799.

    doi: 10.3390/signals4040043

16. Chou CY, Hsu DY, Chou CH. Predicting the onset of diabetes with machine learning methods. *J Pers Med*. 2023;13(3):406.

    doi: 10.3390/jpm13030406

17. Abraham A, Jose R, Ahmad J, *et al*. Comparative analysis of machine learning models for image detection of colonic polyps vs. Resected polyps. *J Imaging*. 2023;9(10):215.

    doi: 10.3390/jimaging9100215

18. Jose R, Syed F, Thomas A, Toma M. Cardiovascular health management in diabetic patients with machine-learning-driven predictions and interventions. *Appl Sci*. 2024;14(5):2132.

    doi: 10.3390/app14052132

19. Jose R, Thomas A, Guo J, Steinberg R, Toma M. Evaluating machine learning models for prediction of coronary artery disease. *Glob Transl Med*. 2024;3(1):2669.

    doi: 10.36922/gtm.2669

20. Freitas ME, Ruiz-Lopez M, Dalmau J, *et al*. Seizures and movement disorders: Phenomenology, diagnostic challenges and therapeutic approaches. *J Neurol Neurosurg Psychiatry*. 2019;90(8):920-928.

    doi: 10.1136/jnnp-2018-320039

21. Song W, Latham NK, Liu L, *et al*. Improved accuracy and efficiency of primary care fall risk screening of older adults using a machine learning approach. *J Am Geriatr Soc*. 2024;72:1145-1154.

    doi: 10.1111/jgs.18776

22. Durán-Vega LA, Santana-Mancilla PC, Buenrostro-Mariscal R, *et al*. An IoT system for remote health monitoring in elderly adults through a wearable device and mobile application. *Geriatrics* (*Basel*). 2019;4(2):34.

    doi: 10.3390/geriatrics4020034

23. Oliver D, Healey F, Haines TP. Preventing falls and fall-related injuries in hospitals. *Clin Geriatr Med*. 2010;26(4):645-692.

    doi: 10.1016/j.cger.2010.06.005

24. Majidi SA, Fakoorfard Z, Safarmohammadi H, Kazemnezhad Leily E. The relationship between moral intelligence and patient safety culture in nurses. *J Caring Sci*. 2023;12: 241-247.

    doi: 10.34172/jcs.2023.30501

25. Hiyama A. Using the analytic hierarchy process to measure nurses' decision-making regarding fall risks and care strategies for fall prevention. *J Nurs Meas*. 2024;31: 1113-1125.

    doi: 10.1891/JNM-2023-0006

26. Orts-Cortés MI, Cabañero-Martínez MJ, Meseguer-Liza C, Arredondo-González CP, de la Cuesta-Benjumea C, Abad-Corpa E. Effectiveness of nursing interventions in the prevention of falls in older adults in the community and in health care settings: A systematic review and meta-analysis of RCT. *Enferm Clín* (*Engl Ed*). 2024;34:4-13.

    doi: 10.1016/j.enfcle.2024.01.001

27. Abbe JR, O'Keeffe C. Continuous video monitoring: Implementation strategies for safe patient care and identified best practices. *J Nurs Care Qual*. 2021;36(2):137-142.

    doi: 10.1097/NCQ.0000000000000502

28. Camp K, Murphy S, Pate B. Integrating fall prevention strategies into EMS services to reduce falls and associated healthcare costs for older adults. *Clin Interv Aging*. 2024;19:561-569.

    doi: 10.2147/CIA.S453961

29. Wabe N, Meulenbroeks I, Huang G, *et al*. Development and internal validation of a dynamic fall risk prediction and monitoring tool in aged care using routinely collected electronic health data: A landmarking approach. *J Am Med Inform Assoc*. 2024:ocae058.

    doi: 10.1093/jamia/ocae058

30. Coleman A. Reducing falls among residents of retirement homes: A DNP project. *Nurse Pract*. 2024;49(4):39-47.

    doi: 10.1097/01.NPR.0000000000000161

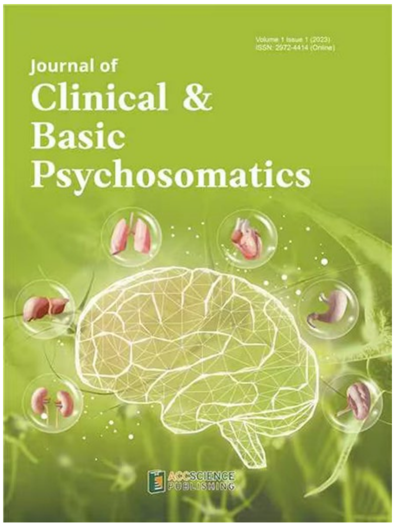
31. Lee-Confer J. Strength in arms: Empowering older adults against the risk of slipping and falling-a theoretical perspective. *Front Sports Act Living*. 2024;6:1371730.

    doi: 10.3389/fspor.2024.1371730

32. Xu D, Wang Y, Zhu S, Zhao M, Wang K. Relationship between fear of falling and quality of life in nursing home residents: The role of activity restriction. *Geriatr Nurs*. 2024;57:45-50.

    doi: 10.1016/j.gerinurse.2024.03.006

Our Journals



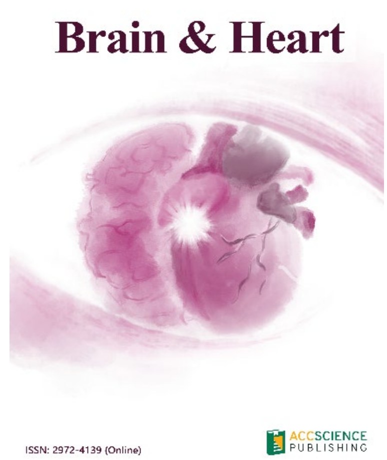
Journal of Clinical and Basic Psychosomatics (JCBP) is a quarterly journal focusing on clinical and basic research on symptoms, assessment, treatment, management, and the mechanism of psychosomatic disorders. *Journal of Clinical and Basic Psychosomatics* covers subject areas, including but not limited to the following:

- Conceptualization and classification of psychosomatic medicine
- Mechanism, biological markers, brain images, and treatment studies
- Psychosomatic reactions, syndromes, disorders, and diseases
- Psychosomatic disorders treated in general hospitals, including endocrinology, neurology, gastroenterology, dermatology, pain management, oncology, rheumatology, and other departments
- Psychological evaluation, management, rehabilitation, resilience training, and psychotherapy for general and specific populations during the pandemic
- Physiological disorders related to psychological factors (eating disorders, sleeping disorders, and sexual dysfunction)
- Somatic symptoms and related disorders and mental disorders due to somatic disease

Brain & Heart focuses on neurocardiology, a neurology and cardiology-based interdisciplinary subject that studies the circulatory mechanism of the human body, as well as the mechanisms of the interplay between the cardiovascular system and the nervous system. The journal's scope includes:

Clinical and basic research on diseases related to the circulatory and nervous systems, such as: orthostatic dizziness, orthostatic hypotension, autonomic dysfunction, and the relationship between the autonomic nervous system and the circulatory function in cerebral degeneration;

Heart-brain research on patients with syncope, autonomic dysfunction, cryptogenic stroke, and stroke with atrial fibrillation; research on the relationship between structural heart diseases and nervous system diseases, the correlation between cardiac electrophysiology and abnormal organizational structures and the pathogenesis of stroke, as well as new ways of diagnosis, treatment and prevention of unexplained stroke.



Start a new journal

Write to us via email if you are interested to start a new journal with AccScience Publishing. Please attach your CV, professional profile page and a brief pitch proposal in your email. We shall inform you of our decision whether we are interested to collaborate in starting a new journal.

Contact: info@accscience.com

https://www.accscience.com/journal/AIH

ACCSCIENCE PUBLISHING

Contact
www.accscience.com
8 Burn Road, #15-03 Trivex, Singapore 369977
Email: editorial@accscience.com
Phone: +65 8182 1586