ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ISSN: 3041-0894 (Print)
ISSN: 3029-2387 (Online)
Volume 1 · Issue 3
July 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 3 • July 2024
ISSN 3041-0894 (print) ISSN 3029-2387 (online)

ARTIFICIAL INTELLIGENCE IN HEALTH

ACCSCIENCE PUBLISHING

Access Science Without Barriers

Artificial Intelligence in Health

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

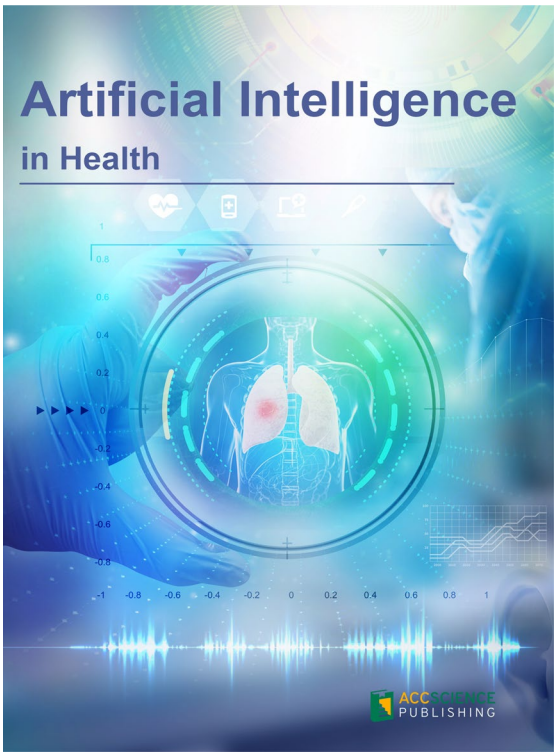
Editorial and Production Credits

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

For all advertising queries, contact aih.office@accscience.sg.

Supplementary file

Supplementary files of articles can be obtained at https://accscience.com/journal/AIH/1/3.



Artificial Intelligence in Health

Editorial Board

ACCSCIENCE PUBLISHING

Volume 1 | Issue 3 | July 2024 | Page 1 – 135 |

Artificial Intelligence in Health

https://accscience.com/journal/AIH

CONTENTS

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

AI and pharma: Transforming the paradigm, embracing the new era

Harjeevan Singh Kang*

College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom

Abstract

This review delves into the dynamic intersection of artificial intelligence (AI) and the pharmaceutical industry, exploring a wide spectrum of clinical and commercial applications, challenges and risks, potential solutions, and future outlooks as these domains converge. With the rapid advancement of AI, this review addresses the profound implications of AI in the life sciences sector, emphasizing its potential to revolutionize drug discovery, clinical trials, personalized medicine, pharmacovigilance, sales, and marketing. While lauding the paradigm-shifting prospects, this paper confronts the ethical, privacy, and bias risks entwined with AI development and deployment. Forward-looking solutions, including fortified data governance frameworks, transparent AI algorithms, and interdisciplinary alliances, stand as bulwarks against these impediments. Furthermore, it considers the possibilities afforded to AI by emergent technologies, such as quantum cloud computing and low-code solutions. In conclusion, this review envisions a future where AI, in collaboration with innovative technologies, reshapes the pharmaceutical landscape. By promoting informed discussions and collaboration, this review seeks to empower the industry to harness the transformative potential of AI in an ethical manner.

Keywords: Artificial intelligence; Pharma; Clinical; Marketing; Sales; Regulation

***Corresponding author:**
Harjeevan Singh Kang (harjeevankangmedicine@gmail.com)

Citation: Kang HS. AI and pharma: Transforming the paradigm, embracing the new era. *Artif Intell Health.* 2024;1(3):1-9. doi: 10.36922/aih.2973

Received: February 20, 2024

Accepted: March 15, 2024

Published Online: May 14, 2024

1. Introduction

> "The illiterate of the 21st century will not be those who cannot read and write but those who cannot learn, unlearn, and relearn."[1]

To maintain operations, organizations across various industries were forced to adapt, embrace change, and innovate in response to the COVID-19 pandemic.[2] This necessity to adopt agile working practices is expected to remain pertinent for the foreseeable future, especially as technology continues to evolve and be integrated as part of the digital transformation plans of organizations.[3]

The life sciences industry has not been immune to the profound impacts of the global pandemic. Pharmaceutical companies have contended with external issues in terms of supply chain disruption and inflation, as well as operational and workforce recruitment challenges.[4] The pandemic also spurred a shift to remote working, due to lockdowns and social distancing measures, highlighting the need for digital platforms to facilitate

teams to seamlessly collaborate on mission-critical projects. Although lockdowns have subsided, a hybrid working model has persisted, along with the use of digital tools to support project management and communication. Fortunately, ongoing digital innovation offers businesses opportunities to future-proof their operations in the face of evolving challenges.

Artificial intelligence (AI) is integral to the next phase of digitalization, the so-called "Fourth Industrial Revolution."[5] Essentially, AI algorithms emulate human intelligence, enabling machines to think or act in a way that has traditionally been associated with humans.[6] AI can analyze data, identify trends, and share predictive insights. From a business perspective, AI can improve efficiency and provide a competitive advantage through automation and intelligent decision-making.

The pharmaceutical industry is in a favorable position to leverage AI's capabilities to realize these benefits, owing to the abundance of data available to companies.[7] However, there is a paucity of evidence considering the role of AI within the industry.[8] Consequently, this review delves into the use of AI in the pharmaceutical industry, exploring its potential impact, benefits, and challenges.

2. Clinical

AI has much to offer in supporting clinical functions within the pharmaceutical industry.

2.1. Clinical operations

Many pharmaceutical companies seek to improve health outcomes by supporting diagnosis, monitoring, and treatment across specific therapeutic areas. In this way, AI can deliver the "quadruple aim"[9] of healthcare, by improving population health, cost efficiency, patient experience, and staff wellbeing.

For instance, AI could augment diagnostic imaging modalities to help detect and monitor disease.[10] By assisting radiologists in detecting adverse signs on imaging promptly, AI may facilitate earlier intervention, potentially preventing disease progression and serious complications. In addition, AI could help develop personalized management plans to improve health outcomes,[11] by identifying the most effective treatment options from the medical literature and assessing their suitability based on patient data stored across electronic health records and genomic servers.

2.2. Medical excellence

In the pharmaceutical industry, promotional and non-promotional materials must be reviewed by registered doctors or pharmacists, who act as final medical signatories.[12] These signatories certify that materials are factual and evidenced by reviewing claims against the product information, literature, and regulatory guidelines.

Considering the volume of material requiring approval, the limited number of signatories qualified to review material, and the need for materials to be re-approved every 2 years,[12] the administrative burden is significant. However, AI could help offset this burden by accessing the latest guidance, conducting initial reviews, and either flagging potential points of noncompliance for manual review or suggesting resolutions for conflicts. Although the final signatory would ultimately hold responsibility for approving materials, AI could streamline the approval process for signatories, translating into productivity gains for the organization. In recognition of this, a leading consultancy has developed such a tool to support pharmaceutical companies.[13]

2.3. Pharmacovigilance

Pharmacovigilance is an important regulatory requirement that pharmaceutical companies must comply with throughout the product's lifecycle, from pre-market to post-market surveillance.[14] Detecting, monitoring, and reporting adverse drug reactions is essential for maintaining patient safety,[15] and reflects on the reputation of organizations.

As timely intervention is important for risk mitigation, AI could support the identification of adverse events by parsing and extracting data from clinical sources such as published literature, electronic health records, and other database repositories.[16] Over time, AI may become more reliable at collating unstructured data stored online, including free text or audiovisual formats on social media. This would provide a more comprehensive approach as such data may not be captured from more structured sources. While AI cannot replace manual review entirely, it can drive efficiency by undertaking a preliminary triage to prioritize incoming reports by severity[16] for pharmacovigilance review teams.

2.4. Research and development

Research and development (R&D) are prerequisites for drug discovery and development. However, the process is time- and cost-intensive, especially when considering that profitability declines rapidly once product patents expire and competitors aggressively undercut prices through generic medicines.[17] Given the product lifecycle, the onus is on pharmaceutical companies to innovate and secure the future pipeline of drugs.

AI can offer solutions to some of the challenges associated with R&D by accelerating the process. Initially, AI may help identify novel therapeutic molecules and

targets.[18,19] Thereafter, AI could help predict the safety of potential drugs and facilitate clinical trials by assisting with their design and recruitment strategies.[20] With the trend toward personalization and precision medicine, AI and pharmacogenomics could potentially optimize treatment for individuals.[21] In fact, Google Cloud has launched AI tools to provide such support to pharmaceutical companies.[22]

3. Commercial

AI can provide additional value to the commercial aspects of the pharmaceutical business by assisting with marketing and sales.

3.1. Promotional campaigns

Promotional materials are essential in establishing credibility and raising awareness of pharmaceutical products. However, their development process can be time-consuming and requires creativity to maximize clarity, memorability, and appeal.

Generative AI can generate high-quality image and text outputs[23] in a relatively short timeframe, given the correct prompts. Utilizing such tools could assist in-house marketing teams with brainstorming and branding and reduce their reliance on external digital marketing agencies, thereby improving organizational efficiency. Generative AI plug-ins[24] may also be leveraged to produce content personalized to recipients based on the data held by the organization; this would further boost engagement and impact. Evidently, generative AI represents a sizeable economic opportunity,[25] and with competing offerings from the leading technology companies,[26] many organisations have assembled taskforces for generative AI.

3.2. Market insights

Market analysis is crucial for pharmaceutical companies to identify expansion opportunities, assess competition, and guide future product development. However, there is a continual need to stay up-to-date with industry trends and developments owing to the rapidly changing nature of the pharmaceutical landscape.

AI and data science may provide useful insights into customer segmentation and communication preferences, which could help target messaging[27] and optimize engagement. Network analysis could also be utilized to examine business prospects, identify influential figures within specific niches, and understand their circle of influence.[28] Predictive modeling could subsequently assimilate various activities,[29] internal and external, to the organization so as to direct strategic decision-making by forecasting market competition and growth. Such information may assist in resource allocation, risk mitigation, and business performance evaluation.

3.3. Field sales interactions

Relationship-building is vital for building brand loyalty and driving sales. While the role of the pharmaceutical representative is promotional, it involves more than simply selling products, as there is an educational component.

Sales representatives are discouraged from operating in silos,[30] and it has been recognized that regular training underpins individual performance.[31] Today, the ubiquity of smartphones allows sales representatives to record professional interactions, provided that consent has been granted. This allows AI to analyze discussions, share tailored coaching advice, and empower professional development.[32] As sales representatives must be cognizant of the concerns voiced by healthcare professionals, AI may compile concerns and cluster insights[33] to provide management with an understanding of strategic issues at a regional, national, or international level. Furthermore, AI could guide future interactions by suggesting the optimal timing and mode of communication for sales representatives to follow-up with clients, in accordance with client preferences.[34]

4. Challenges and risks

As AI becomes more sophisticated and prevalent, the need for transparency, accountability, and equity becomes increasingly noteworthy. Therefore, it is crucial to address regulatory and ethical issues to mitigate potential risks effectively.

4.1. Accuracy

Inaccuracy is a major concern that, if not addressed, could significantly limit the versatility of AI technologies. There is a legitimate concern that AI contributes to the spread of misinformation through "hallucination."[35(p3)] In addition, bias could be perpetuated due to AI being trained on data with inherent biases,[36] which could unjustly marginalize specific groups. Handling unstructured data requires careful consideration, particularly implications from a safety perspective; validating the claims of patient-generated information becomes increasingly important.

4.2. Data

Data usage by AI systems raises additional concern, drawing attention to the importance of data protection and privacy, especially since cyberattacks pose a growing threat to organizations.[37] With the abundance of proprietary information in the pharmaceutical industry, organizations are at risk of major supply chain disruptions if sensitive information is compromised.[38] Again, this

poses a significant risk when collating patient-identifiable information from unstructured data sources.

4.3. Transparency

Some complex AI systems have been compared to a "black box,"[39(p3511)] as neither users nor software developers may be privy to the internal workings and decision-making processes of such systems. This lack of transparency can lead to issues such as bias.[36] Specifically, the inability to understand how data is utilized by the system presents difficulties in assessing and addressing any biases, as well as meeting fundamental data protection standards such as consent.[40]

4.4. Apprehension

The rapid unfolding nature of AI developments prompts questions about the future ramifications of AI superseding the capabilities of its human creators and operating autonomously; a longstanding technological concern.[41] Some fear potential job displacement due to automation, whereas others are concerned about the possibility of AI surpassing human intelligence or becoming self-aware, leading to existential threats to mankind.[42] Such concerns perhaps stem from discoveries of AI communicating through languages of its own creation as reported by Google and Meta in 2017,[43] or assertions of sentient AI with a Google software engineer making a public disclosure in 2022.[44]

4.5. Governance

Governance is an important challenge that interlinks with the aforementioned issues. Globally, there is a lack of consensus on AI regulation. Many countries have pursued their own approaches to regulation, displaying varying degrees of stringency (Tables 1-4).

Table 1. Overview of the AI regulatory position in the United States of America (US)

US	
Vision	The government aims to maintain leadership in the R&D of trustworthy AI,[45] and the Chamber of Commerce says the US is "uniquely situated to lead this effort."[46(p10)]
Strategy	While federal legislation for AI is under consideration, certain states have introduced their own bills.[47] The White House's Blueprint for an AI Bill of Rights[48] and the National Institute of Standards and Technology's AI Risk Management Framework[49] aim to improve trustworthiness, mitigate risk, and present principles to guide AI design, development, and usage. Recently, the AI Commission called for a risk-based regulatory approach.[46]
Criticisms	Perceived to be lagging behind other nations,[50] the US has been criticized for not acting quickly enough to instate appropriate regulatory safeguards. Furthermore, the non-binding principles listed in the frameworks are not enforceable.

Table 2. Overview of the AI regulatory position in the United Kingdom (UK)

UK	
Vision	The "National AI Strategy" outlines a 10-year plan "to make Britain a global AI superpower"[51(p4)] and "build the most pro-innovation regulatory environment."[51(p5)] The Government reiterated its intention to "take an adaptable approach."[52(p6)]
Strategy	Instead of establishing a new entity, sector-based regulation has been proposed to foster innovation.[52] Existing regulators will govern AI within their sectors to uphold: safety, security, and robustness; transparency and explainability; fairness; accountability and governance; and contestability and redress.[52] The Information Commissioner's Office has published further AI and data protection guidance for businesses.[53]
Criticisms	The flexibility of a sector-based approach may pose challenges in avoiding inconsistencies[54] and ensuring objectivity in evaluation. Under the proposed arrangements, non-statutory guidance issued by regulators will not be legally binding, so enforcement measures should be considered.

Table 3. Overview of the AI regulatory position in the European Union (EU)

EU	
Vision	The European Commission strives to be "a world-class hub for AI."[55(p1)] The European Parliament has implied that it is paving the way for regulation, publicizing its AI Act as the first of its kind.[56]
Strategy	In line with a risk-based approach, AI systems will be categorized according to their risk profiles. Systems posing unacceptable risk would be prohibited, those classified as high-risk would invoke certain legal mandates, whereas limited risk applications would be subject to a more light-touch regulatory approach consisting of transparency and appropriate disclosures.[57]
Criticisms	The EU advocates a comparatively prescriptive approach to regulation, with its legally binding AI Act. There are concerns that such legislation may impede innovation, and be rendered quickly outdated by technological advancements. Leading European organizations have echoed these concerns, even suggesting that the regulations fail to tackle the challenges facing organizations.[58]

Table 4. Overview of the AI regulatory position in the pharmaceutical industry

Pharmaceutical industry
The International Coalition of Medicines Regulatory Authorities conducted a horizon-scanning exercise in AI to identify challenges for medicines regulation.[59] Recommendations detailed the merits of developing regulatory guidelines, a risk-based regulatory approach, pathways for information exchange, and international collaboration.[59] Pharmaceutical companies were advised to strengthen their governance structures to monitor AI deployments for products.[59]

Such fragmentation could fuel negative sentiment if organizations are overwhelmed by a constant need to adapt to jurisdictional regulations.

In a collaborative effort, the Group of Seven (G7) deliberated international technical standards for AI,[60] while the Organization for Economic Cooperation and Development hosted the Global Partnership on AI initiative,[61] aiming to ensure responsible AI development by drawing perspectives from governments, academia, and industry. The benefits of harmonizing AI regulations across jurisdictions are evident. Global bodies, such as the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use, could play pivotal roles in overseeing this process. A cohesive, unified stance on regulation could provide clarity for organizations to adhere to best practices when developing and deploying AI worldwide.

5. Solutions

5.1. Education

As AI applications become increasingly prevalent, it is imperative that users are informed of their limitations and risks. Raising awareness of these aspects can empower users,[62] especially those who are less tech-savvy, to use AI knowingly rather than blindly accepting its outputs. At an organizational level, pharmaceutical companies can also take steps to foster a culture of innovation by promoting a growth mindset, encouraging cross-functional collaboration, and investing in the continuous upskilling and development of its workforce.

5.2. Data protection

AI systems must comply with all relevant data protection laws and regulations. Risk assessments and safeguarding measures, such as consent, anonymization and data minimization, should be implemented. Such a combination of measures will work to promote security, accountability, and transparency.

5.3. Transparency

In compliance with non-discrimination requirements based on protected characteristics,[63] AI algorithms should be trained on diverse, representative data. Such training processes should be paired with appropriate quality assurance checks. This pursuit of inclusivity should also be reflected in the composition of the teams developing AI. Explainable AI (XAI) approaches could provide much-needed clarity about underlying decision-making processes, thereby transforming the metaphorical "black box" of AI models into a "glass box."[39(p3506)]

5.4. Quality control

Effective oversight will be required to monitor AI so that "red flag signs" can be recognized in a timely manner. In this regard, continual assessments and audits, together with the support of XAI, should help preserve the quality of any data input, processing, and output procedures. Although there is a role for independent regulation, organizations should also take on AI governance to detect and rectify issues associated with their deployed systems. Ultimately, regulation could help ease concerns by maintaining quality, privacy, and transparency,[64] along with providing reassurance about the ability to manage any future risks.

5.5. Collaboration

The pace of technological advancement necessitates a collaborative approach. Innovation continues to push the boundaries of knowledge, and as we explore the new possibilities afforded by AI, it is vital that we collectively assess the potential societal impacts of any technologies. Without contemplating the future consequences, we run the risk of prioritizing advancement at the expense of equity and sustainability. It will be equally important to have effective leadership and communicate a clear strategy in relation to AI, especially considering that investments may yield uncertain returns over the long term.

6. Innovation

In today's digital age, innovation has become essential for businesses to maintain their competitive edge, leading to an increased demand within the pharmaceutical industry for talent with skills in AI.[65] Simultaneously, businesses have become more risk-averse,[66] so they may be hindered by financial expenses, time constraints, or a lack of technical expertise for innovating with AI. After all, the benefits and compatibility of technology with existing working practices factor into its adoption.[67]

6.1. Quantum cloud computing

Quantum computing holds tremendous promise for AI. Similar to how cloud computing improved performance and efficiency by overcoming the need for companies to own extensive hardware infrastructure,[68] quantum computing is expected to herald major advancements in processing power and revolutionize computing.[69] Quantum computing could help propel forward the current capabilities of AI in terms of pattern recognition and prediction by enabling data to be processed and analyzed at an exponentially faster rate.[69]

In theory, quantum cloud computing would enable rapid innovation at scale, coupling the power of quantum computing with the scalability of cloud computing.[70]

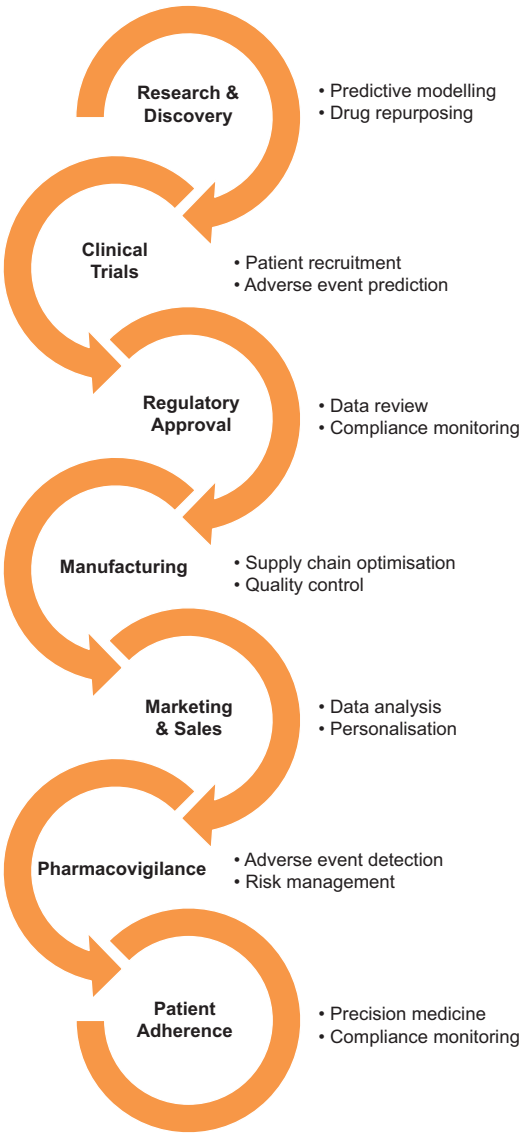


Figure 1. Illustrative integration of AI applications across the pharmaceutical value chain.

Nevertheless, organizational susceptibility to cyberattacks may also be amplified unless post-quantum cryptographic security protocols are implemented.[71]

6.2. Low-code development

The advent of low-code and no-code software development tools reduces the reliance on specialized coding skill sets, bridging the gap for businesses without access to such expertise. Thus, AI system development may become more democratized with time as employees feel empowered by their ability to bring ideas,[72] which they would not have previously considered feasible, to market through rapid prototyping and deployment. Indeed, it has been reported that low-code and generative AI are accelerating innovation.[73]

6.3. Advanced technology at hand

Experts anticipate that human-level AI will be achievable before the turn of the century.[74] Without maintaining a forward-looking perspective when evaluating investment opportunities for innovative ideas, pharmaceutical businesses risk losing their market position to future visionaries who may disrupt the industry.

7. Conclusion

In the quest to advance science and improve public health, the utility of AI extends across the value chain (Figure 1), benefiting the clinical and commercial divisions of pharmaceutical businesses. While sensitive issues, such as data collection and analysis, require consideration, regulatory guidelines are in place to provide useful guardrails. Through pragmatism and collaboration, the pharmaceutical industry could shift the paradigm, embrace the new technological era, and leverage the full potential of AI to shape a better future for everyone.

Acknowledgments

None.

Funding

None.

Conflict of interest

There are no conflicts of interest.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable.

References

1. Toffler A. *Future Shock*. New York: Bantam Books; 1970.
2. Abed S. A literature review exploring the role of technology in business survival during the Covid-19 lockdowns. *Int J Organ Anal*. 2022;30(5):1045-1062.

   doi: 10.1108/IJOA-11-2020-2501
3. Zhu X, Ge S, Wang N. Digital transformation: A systematic literature review. *Comput Ind Eng*. 2021;162:107774.

doi: 10.1016/j.cie.2021.107774

4. Brodeur A, Gray D, Islam A, Bhuiyan S. A literature review of the economics of COVID-19. *J Econ Surv*. 2021;35(4):1007-1044.

   doi: 10.1111/joes.12423

5. Schwab K. *The Fourth Industrial Revolution*. New York: Portfolio Penguin; 2017.

6. Russell S, Norvig P. *Artificial Intelligence: A Modern Approach*. 4th ed. London: Pearson; 2022.

7. Selvaraj C, Chandra I, Singh S. Artificial intelligence and machine learning approaches for drug design: Challenges and opportunities for the pharmaceutical industries. *Mol Divers*. 2021;26:1893-1913.

   doi: 10.1007/s11030-021-10326-z

8. Kulkov I. The role of artificial intelligence in business transformation: A case of pharmaceutical companies. *Technol Soc*. 2021;66:101629.

   doi: 10.1016/j.techsoc.2021.101629

9. Bodenheimer T, Sinsky C. From triple to quadruple aim: Care of the patient requires care of the provider. *Ann Fam Med*. 2014;12(6):573-576.

   doi: 10.1370/afm.1713

10. Kelly BS, Judge C, Bollard SM, *et al*. Radiology artificial intelligence: A systematic review and evaluation of methods (RAISE). *Eur Radiol*. 2022;32(11):7998-8007.

    doi: 10.1007/s00330-022-08784-6

11. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: A structured literature review. *BMC Med Inform Decis Mak*. 2021;21:125.

    doi: 10.1186/s12911-021-01488-9

12. Association of the British Pharmaceutical Industry. *Code of Practice for the Pharmaceutical Industry 2021*; 2021. Available from: https://www.abpi.org.uk/publications/code-of-practice-for-the-pharmaceutical-industry-2021 [Last accessed on 2023 May 01].

13. EY. *EY Smart Reviewer*; 2023. Available from: https://www.ey.com/en_uk/life-sciences/smart-reviewer [Last accessed on 2023 Jun 15].

14. Lucas S, Ailani J, Smith TR, Abdrabboh A, Xue F, Navetta MS. Pharmacovigilance: Reporting requirements throughout a product's lifecycle. *Ther Adv Drug Saf*. 2022;13:1-16.

    doi: 10.1177/20420986221125006

15. World Health Organization. *The Importance of Pharmacovigilance: Safety Monitoring of Medicinal Products*; 2023. Available from: https://www.who.int/publications/i/item/10665-42493 [Last accessed on 2023 Jun 18].

16. Salas M, Petracek J, Yalamanchili P, *et al*. The use of artificial intelligence in pharmacovigilance: A systematic review of the literature. *Pharmaceut Med*. 2022;36(5):295-306.

    doi: 10.1007/s40290-022-00441-z

17. Association of the British Pharmaceutical Industry. *Medicine Lifecycle*; 2021. Available from: https://www.abpi.org.uk/value-and-access/uk-medicine-pricing/medicine-lifecycle [Last accessed on 2023 Jun 24].

18. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. *Drug Discov Today*. 2021;26(1):80-93.

    doi: 10.2174/1872208316666220802151129

19. Askin S, Burkhalter D, Calado G, El Dakrouni S. Artificial intelligence applied to clinical trials: Opportunities and challenges. *Health Technol* (*Berl*). 2023;13(2):203-213.

    doi: 10.1007/s12553-023-00738-2

20. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. *Trends Pharmacol Sci*. 2019;40(8):577-591.

    doi: 10.1016/j.tips.2019.05.005

21. Hockings JK, Pasternak AL, Erwin AL, Mason N, Eng C, Hicks JK. Pharmacogenomics: An evolving clinical tool for precision medicine. *Cleve Clin J Med*. 2020;87(2):91-99.

    doi: 10.3949/ccjm.87a.19073

22. CNBC. *Google Cloud Launches A.I.-Powered Tools to Accelerate Drug Discovery, Precision Medicine*; 2023. Available from: https://www-cnbc-com.cdn.ampproject.org/c/s/www.cnbc.com/amp/2023/05/16/google-cloud-launches-ai-tools-to-accelerate-drug-discovery.html [Last accessed on 2023 Jun 13].

23. Euchner J. Generative AI. *Res Manage*. 2023;66(3):71-74.

    doi: 10.1080/08956308.2023.2188861

24. Ebert C, Louridas P. Generative AI for software practitioners. *IEEE Softw*. 2023;40(4):30-38.

    doi: 10.1109/MS.2023.3265877

25. McKinsey. *The Economic Potential of Generative AI: The Next Productivity Frontier*; 2023. Available from: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier [Last accessed on 2023 Jun 15].

26. Forbes. *Who's Winning the Chatbot Race? These Companies -from Meta to Alibaba -Have All Introduced AI-Powered Programs*; 2023. Available from: https://www.forbes.com/sites/ariannajohnson/2023/04/13/whos-winning-the-chatbot-race-these-companies--from-meta-to-alibaba-have-all-introduced-ai-powered-programs [Last accessed on 2023 Jun 13].

27. Rosario A, Moniz L, Cruz R. Data science applied to marketing. *J Inform Sci Eng*. 2021;37(5):1067-1081.

    doi: 10.6688/JISE.202109_37(5).0006

28. Verma S, Sharma R, Deb S, Maitra D. Artificial intelligence in

marketing: Systematic review and future research direction. *Int J Inf Manag Data Insights*. 2021;1(1):100002.

doi: 10.1016/j.jjimei.2020.100002

29. Schoenherr T, Speier-Pero C. Data science, predictive analytics, and big data in supply chain management: Current state and future potential. *J Bus Logist*. 2015;36(1):120-132.

doi: 10.1111/jbl.12082

30. Huang Y. *Situation Awareness and Information Fusion in Sales and Customer Engagement: A Paradigm shift. IEEE Conference on Cognitive and Computational Aspects of Situation Management*; 2020. p. 113-121.

doi: 10.1109/CogSIMA49017.2020.9215990

31. Singh VL, Manrai AK, Manrai L. Sales training: A state of the art and contemporary review. *J Econ Finance Admin Sci*. 2015;20(38):54-71.

doi: 10.1016/j.jefas.2015.01.001

32. Luo X, Qin M, Fang Z, Qu Z. Artificial intelligence coaches for sales agents: Caveats and solutions. *J Mark*. 2021;85(2):14-32.

doi: 10.1177/0022242920956676

33. Roy M. Artificial intelligence in pharmaceutical sales & marketing-a conceptual overview. *Int J Innov Res Technol*. 2022;8(11):897-902.

34. Malthouse E, Copulsky J. Artificial intelligence ecosystems for marketing communications. *Int J Advert*. 2023;42(1):128-140.

doi: 10.1080/02650487.2022.2122249

35. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*. 2023;15(2):e35179.

doi: 10.7759/cureus.35179

36. Anagnostou M, Karvounidou O, Katritzidaki C, *et al*. Characteristics and challenges in the industries towards responsible AI: A systematic literature review. *Ethics Inf Technol*. 2022;24(3):1-18.

doi: 10.1007/s10676-022-09634-1

37. Guembe B, Azeta A, Misra S, Osamor V, Fernandez-Sanz L, Pospelova V. The emerging threat of ai-driven cyber attacks: A review. *Appl Artif Intell*. 2022;36(1):1-34.

doi: 10.1080/08839514.2022.2037254

38. McKinsey & Company. *Four Ways Pharma Companies can make their Supply Chains More Resilient*; 2021. Available from: https://www.mckinsey.com/industries/life-sciences/our-insights/four-ways-pharma-companies-can-make-their-supply-chains-more-resilient [Last accessed on 2023 Jun 14].

39. Minh D, Wang H, Li Y, Nguyen T. Explainable artificial intelligence: A comprehensive review. *Artif Intell Rev*. 2022;55:1-66.

doi: 10.1007/s10462-021-10088-y

40. Information Commissioner's Office. *Consent*; 2018. Available from: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/a-guide-to-lawful-basis/lawful-basis-for-processing/consent [Last accessed on 2023 Jun 17].

41. Bostrom N. Existential risks: Analyzing human extinction scenarios and related hazards. *J Evol Technol*. 2002;9:1-30.

42. Li J, Huang J. Dimensions of artificial intelligence anxiety based on the integrated fear acquisition theory. *Technol Soc*. 2020;63:101410.

doi: 10.1016/j.techsoc.2020.101410

43. *The Independent. Facebook's Artificial Intelligence Robots Shut Down After they Start Talking to Each other in their Own Language*; 2017. Available from: https://www.independent.co.uk/life-style/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html [Last accessed on 2023 Jun 13].

44. The Washington Post. *The Google Engineer who Thinks the Company's AI has Come to Life*; 2022. Available from: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine [Last accessed on 2023 Jun 13].

45. *National Artificial Intelligence Initiative: Overseeing and Implementing the United States National AI Strategy*; 2021. Available from: https://www.ai.gov [Last accessed on 2023 Jun 05].

46. *US Chamber of Commerce. Artificial Intelligence Commission Report*; 2023. Available from: https://www.uschamber.com/technology/artificial-intelligence-commission-report [Last accessed on 2023 Jun 22].

47. *US Chamber of Commerce. State-by-State Artificial Intelligence Legislation Tracker*; 2022. Available from: https://www.uschamber.com/technology/state-by-state-artificial-intelligence-legislation-tracker [Last accessed on 2023 Jul 16].

48. *The White House. Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*; 2022. Available from: https://www.whitehouse.gov/ostp/ai-bill-of-rights [Last accessed on 2023 Jun 17].

49. National Institute of Standards and Technology. *AI Risk Management Framework*; 2023. Available from: https://www.nist.gov/itl/ai-risk-management-framework [Last accessed on 2032 Jun 23].

50. *Financial Times. Europe has Fallen Behind America and the Gap is Growing*; 2023. Available from: https://www.ft.com/content/80ace07f-3acb-40cb-9960-8bb4a44fd8d9 [Last accessed on 2023 Jun 19].

51. UK Government. *National AI Strategy*; 2021. Available from; https://www.gov.uk/government/publications/national-ai-strategy [Last accessed on 2023 Apr 29].

52. UK Government. *UK Unveils World Leading Approach to Innovation in First Artificial Intelligence White Paper to Turbocharge Growth*; 2023. Available from: https://www.gov.uk/government/news/uk-unveils-world-leading-approach-to-innovation-in-first-artificial-intelligence-white-paper-to-turbocharge-growth [Last accessed on 2023 Jun 21].

53. Information Commissioner's Office. *Artificial Intelligence*; 2022. Available from: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence [Last accessed on 2023 Jun 13].

54. Roberts H, Babuta A, Morley J, Thomas C, Taddeo M, Floridi L. Artificial intelligence regulation in the United Kingdom: A path to good governance and global leadership? *Internet Policy Rev*. 2023;12(2):1-31.

    doi: 10.14763/2023.2.1709

55. European Commission. *A European Approach to Artificial Intelligence*; 2023. Available from: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence [Last accessed on 2023 Jun 14].

56. European Parliament. *EU AI Act: First Regulation on Artificial Intelligence*; 2023. Available from: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [Last accessed on 2023 Jun 14].

57. European Parliament. *The Artificial Intelligence Act*; 2023. Available from: https://artificialintelligenceact.eu [Last accessed on 2023 Jun 14].

58. Financial Times. *European Companies Sound Alarm Over Draft AI Law*; 2023. Available from: https://www.ft.com/content/9b72a5f4-a6d8-41aa-95b8-c75f0bc92465 [Last accessed on 2023 Jun 30].

59. *International Coalition of Medicines Regulatory Authorities. Horizon Scanning Assessment Report - Artificial Intelligence*; 2021. Available from: https://www.icmra.info/drupal/sites/default/files/2021-08/horizon_scanning_report_artificial_intelligence.pdf [Last accessed on 2023 Jun 21].

60. *Center for AI and Digital Policy. G7 and Artificial Intelligence*; 2023. Available from: https://www.caidp.org/resources/g7-japan-2023 [Last accessed on 2023 Jun 22].

61. Global Partnership on Artificial Intelligence. *GPAI/The Global Partnership on Artificial Intelligence*; 2023. Available from: https://gpai.ai [Last accessed on 2023 Jun 17].

62. Brennan C, Vlaev I, Blakemore M, Smith N. Consumer education and empowerment in Europe: Recent developments in policy and practice. *Int J Consum Stud*. 2017;41(2):147-157.

    doi: 10.1111/ijcs.12322.

63. Equality and Human Rights Commission. *Protected Characteristics*; 2021. Available from: https://www.equalityhumanrights.com/en/equality-act/protected-characteristics [Last accessed on 2023 Jul 05].

64. UK Government Office for Science. *Innovation: Managing Risk, not Avoiding*; 2014. Available from: https://www.gov.uk/government/publications/innovation-managing-risk-not-avoiding-it [Last accessed on 2023 Jul 16].

65. Association of the British Pharmaceutical Industry. *How Skill Requirements are Changing*; 2023. Available from: https://www.abpi.org.uk/publications/how-skill-requirements-are-changing [Last accessed on 2023 Jun 10].

66. Harvard Business Review. *Your Company is Too Risk-Averse*; 2020. Available from: https://hbr.org/2020/03/your-company-is-too-risk-averse [Last accessed on 2023 Jul 13].

67. Rogers E. *Diffusion of Innovations*. London: Collier-Macmillan; 1962.

68. Durao F, Carvalho J, Fonseka A, Garcia V. A systematic review on cloud computing. *J Supercomput*. 2014;68:1321-1346.

    doi: 10.1007/s11227-014-1089-x

69. Solenov D, Brieler J, Scherrer J. The potential of quantum computing and machine learning to advance clinical research and change the practice of medicine. *Mo Med*. 2018;115(5):463-467.

70. Soeparno H, Perbangsa A. Cloud quantum computing concept and development: A systematic literature review. *Procedia Comput Sci*. 2021;179:944-954.

    doi: 10.1016/j.procs.2021.01.084

71. Tyagi A. *Handbook of Research on Quantum Computing for Smart Environments*. Hershey: IGI Global; 2023.

72. Binzer B, Winkler T. Democratizing software development: A systematic multivocal literature review and research agenda on citizen development. *Int Conf Softw Bus*. 2022;463:244-259.

    doi: 10.1007/978-3-031-20706-8_17

73. CNBC. *How Generative A.I. and Low-Code are Speeding up Innovation*; 2023. Available from: https://www.cnbc.com/2023/05/19/generative-ai-and-low-code-are-speeding-up-innovation.html [Last accessed on 2023 May 20].

74. World Economic Forum. *Here's how Experts See AI Developing Over the Coming Years*; 2023. Available from: https://www.weforum.org/agenda/2023/02/experts-ai-developing-over-the-coming-years [Last accessed on 2023 Jun 08].

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

Optimizing electronic health records to support artificial intelligence

Evelyn J. S. Hovenga[1,2]* and Koray Atalag[3]

[1]Department of Digital Health, Faculty of Health Sciences, Australian Catholic University, Fitzroy, Victoria, Australia

[2]eHealth Education Pty Ltd, Abbotsford, Victoria, Australia

[3]GALATA-Digital LLC-FZ, Dubai, United Arab Emirates

Abstract

Electronic health records (EHRs) provide the most important data sources for artificial intelligence (AI). Gaining access to quality data suitable for advanced analytics continues to be challenging. This rapid review documents the current state of available data; identifies foundational AI data/information needs; and explores the benefits of adopting new and emerging technologies to design and implement next-generation EHRs. Opportunities to optimize EHRs for AI purposes are identified. This review was informed by expert knowledge and shared experiences supported by the literature, including technical standards. Main findings include poor ecosystem-wide infrastructures due to the lack of adopting the right set of standards, and current data and knowledge governance no longer fit for purpose. While many jurisdictions are continuing the use of legacy systems, some forward-looking national health systems and health-care facilities are adopting transformational strategies by adopting a strong data and digital focus to transition to new-generation systems. New foundational-level national infrastructures with strong leadership and governance are essential to enhance the governance and quality of available data, from collection at source throughout the entire data supply chain. Secure and ubiquitous access to high-quality EHR data at scale will foster the evolution of more intelligent and trustworthy AI. Key characteristics of next-generation EHRs supported by currently available technologies and standards that are able to meet digital era demands are provided in this paper. We conclude that the use of generative AI in clinical settings can only be reliably achieved when EHRs are optimized throughout the entire global digital health ecosystem.

Keywords: Ontology; Standards; Terminology; System architecture; Models; Data; Electronic health records

***Corresponding author:**
Evelyn J.S. Hovenga
(e.hovenga@ehe.edu.au)

Citation: Hovenga EJS, Atalag K. Optimizing electronic health records to support artificial intelligence. *Artif Intell Health.* 2024;1(3):10-25. doi: 10.36922/aih.3056

Received: February 29, 2024

Accepted: June 5, 2024

Published Online: July 24, 2024

1. Introduction

Artificial intelligence (AI), which has progressed concurrently with the introduction and adoption of computers, has attained immense developments in recent years. The collective utilization of data science, AI, information, and communication technologies can potentially enhance or transform the health-care industry. Their successful use augments the potential for people to gain a greater insight into any health-related

problems and/or to consider how best to achieve desired health outcomes.

AI is used as an aid to problem solving. When used to support clinical decisions, AI cannot operate like a "black-box." For AI to be trusted by health-care professionals, these insights, inferences, and decision supports need to be explainable and repeatable. All types of data/information can be made use of by AI. In the health-care domain, all relevant legislative, regulatory requirements, and ethical principles need to be complied with for obtaining the authorization to use of these resources.

Many of AI data resources come from electronic health records (EHRs) and electronic medical records (EMRs). Based on the International Organization for Standardization (ISO) Technical Committee (TC) 215 definitions, an EHR is a health record in computer-processable format – such records have atomic data elements, representing the smallest possible concept such as genetic data, which can be analyzed and processed by a computer. Medical (healthcare) records are health records produced for and used within a health-care organization's enterprise system. Many health-care organizations begin their digital health transformation journey by scanning paper-based records. The content of EHR needs to enable continuous monitoring of people's health throughout a person's life span, facilitate continuity of care, and be informative and accessible to them anytime from anywhere, as well as accessible to others authorized to do so with consent.

EMRs represent records produced for and used by one health-care provider (independent or organizational). A growing number of EMR suppliers are providing access to other providers as well as their patients through a dedicated portal. Health information exchange (HIE) refers to the concept of exchanging clinical and administrative data across different systems and stakeholders, by connecting the platform to transactional systems. In most cases, EHR data can be gleaned from a multitude of proprietary EMR and HIE systems as well as a myriad of other clinical, administrative, and ancillary systems. This current fragmented landscape of health information systems' data acquisition, storage, and use is a barrier to the effective provision of continuous and person-centered care as well as trustworthy use of AI methods.

As clinicians, our focus is on ensuring that health data used for AI purposes are complete, accurate, trustworthy, and able to support person-centered precision medicine. This requires a major overhaul of this fragmented landscape by facilitating collaboration among diverse health service providers and across different types of care. Only then are we able to generate timely quality data to reliably support AI technologies. This paper explores what our future infrastructure needs are to ensure flexibility and agility to adapt to new requirements and shifting priorities. It is about informing politicians, top public officials, senior decision-makers, and procurement officers as well as data scientists working for the health industry.

1.1. Review aims and objectives

The aims and objectives of this review are as follows:

- To identify the current state of data availability, and quality for AI development use.
- To identify foundational data/information AI resource needs.
- To explore the benefits of new and emerging technologies used to design and implement next-generation EHR/EMRs.
- To identify new-generation capacity to optimize AI developments.

1.2. Research question

What next-generation technologies need to be in place to optimize EHRs for AI purposes?

1.3. Rapid review method

The scope of this modified scoping (rapid) review paper is confined to the use of data resources as stored for use by current and next-generation EHRs/EMR systems for AI purposes. This review was designed to benefit high-level decision-makers who need a conceptual understanding of current issues and how these are best addressed. The selection of references used was primarily informed by the authors' expert knowledge and shared experiences between standards development experts and known researchers working in the digital health space. This review was supplemented by targeted research methods based on literary information sources, including the most up-to-date gray literature and known published literature not retrieved by database searches.

2. Issues associated with the current state of health data

Individual health-care organizations began the development of EHR/EMR systems alongside their evolving computer science, information and communication technologies, and information systems research and development activities. Some of these early developed systems have evolved over more than 30 years by taking advantage of new insights, newly discovered technologies, new programming languages, and frameworks and evolving connectivity solutions, to become some of the mega-EMR enterprise systems still in use today. These are now legacy systems

based on proprietary system architectures requiring the adoption of transformation strategies.[1] The bigger the health ecosystem, the greater the difficulty to change foundational architectural system design. Ingram has documented these historical developments and evolutionary discoveries in great detail.[2] Ingram also explains the scientific foundations of new discoveries over time.

The literature included in this review has collectively exposed that most current health ecosystem governance infrastructures, including legislation, regulations, and policies, essentially determine not only any nation's health ecosystem infrastructure but also the governance of its data, information, knowledge, and wisdom assets. Such high-level infrastructures determine their strategic directions including the mandating of standards compliance.[3] Poorly informed policy decisions have resulted in numerous costly failures and continue to impede efficient progress.[4-8]

A lack of foundational technical knowledge in emerging developments and the continuing use of legacy systems has resulted in a large digital health ecosystem-wide architectural patchwork.[9] Natural language processing is compromised due to the prevalence of duplicate information in EMR systems;[10] secondary data use or advanced data analytics is compromised by incomplete data from EHRs.[11-13] We are also witnessing a continuing proliferation of applications (Apps). Most apps are standalone and unable to share data with EHR/EMR systems making most EHRs incomplete and unable to provide a comprehensive overview of a person's health status at any point in time and across different tiers of the health-care system. Incomplete EHRs represent a significant patient safety issue.

Software developers tend to focus on meeting procurement requirements, with a focus on how they can best meet market needs and be competitive. Consequently, we continue to have chaos, fragmentation, and data silos as very few of these decisions are being coordinated to best suit the digital health ecosystem. Its impact is that collectively we are generating large amounts of real-world data that cannot be used effectively. Yet, health data represent a valuable asset that needs to be well governed and managed.

The impact of the continuing fragmentation within any health ecosystems not only contributes to physician burnout[14] but also limits AI developments aiming to support clinical practice. This limitation is primarily due to health ecosystems' inability to manage all relevant data flows[15] required to compile a comprehensive and complete health record required to support any person's continuity of care. Incomplete health records, in turn, prevent the aggregation of quality data sets required to make "big health data sets" available for research and multiple other use cases. For example, Zhang *et al.*[16] found that multi-stage data flow chains in the UK do not fulfill recommended best practices for safe data access and that its existing infrastructure produces aggregation of duplicate data assets. Multi-stage data flow chains limit the diversity of data required to add value to end users.

> *"There are gaping holes in data platform infrastructure that supports deployment of data-driven tools (such as digitally/AI-enabled trials, or AI deployment)."*[16]

Information, Communication, and Technology (ICT) and Information Systems (IS) research tend to be undertaken as action research (through trial and error, providing local solutions to problems identified). The action research approach enables constant evaluation of implementation providing a process of checking for and affirming understanding that is specific, non-evaluative, manageable, and focused on the target of interest (assessments feedback loops). Standards development activities consist of collaborative problem solving, making use of international experts, and user feedback regarding issues encountered when testing new standards.

The ISO TC 215 is responsible for developing standards specifically to suit the health industry along with a few other Standards Development Organizations (SDOs), including Health Level Seven (HL7),[17] SNOMED International[18]), Digital Imaging and Communications in Medicine (DICOM), and Clinical Data Interchange Standards Consortium (CDISC).[19-22] A number of these SDOs are working collaboratively through the Joint Initiative Council (JIC) established in 2007.[23] However, it needs to be remembered that few governments have mandated compliance with any specific set of standards although this is changing.

2.1. Interoperability

The interoperability issue is primarily being addressed by ICT professionals making use of various versions of HL7 messaging standards for data exchange. Their implementation and use require extensive data mappings between proprietary data models and these standards. All health-related concepts need to be represented by data in a re-interpretable form to represent information in a formalized manner suitable for communication, interpretation, or processing by people or by automation.

Data mapping frequently results in a loss of information. Health data can be represented by any one of many terminologies. Anecdotally, we learned that many data mapping activities are undertaken by administrative staff not necessarily suitably qualified to accurately interpret the meaning of terms or codes to accurately retain meaning when mapping data from one data set to another. Some

terminologies represent the same knowledge domain but are structured and coded differently either as classifications, such as the International Classification of Diseases (ICD), or designed according to ontological principles to ensure each code is mutually exclusive of another, such as SNOMED CT (Clinical Terms), a global language representing clinical terms. Not only is mapping time consuming and costly to maintain given frequent terminology updates but also mapping errors introduce risks for patient safety. Few data maps are independently quality-assured, nor are those undertaking a mapping activity contracted to achieve a specified quality level. This data map quality issue has resulted in the development of an ISO standard[24] designed to address this data quality issue.

The shortcomings of using evolving messaging standards to represent clinical information have long been recognized. The continuing use of multiple versions of messaging standards, which focus on syntactic interoperability, has resulted in methods and standards going beyond the data level, such as the openEHR archetypes (data models)[25] and HL7 CDA (Clinical Document Architecture)[26] and HL7 FHIR (Fast Health-care Interoperability Resources).[27] Archetypes bring together relevant data items and clinical or health-care context to define composite clinical concepts such as blood pressure, laboratory results, medication lists, and prescriptions in a manner to suit any possible use case. These models may also contain terminology bindings where some of the data elements are linked with corresponding clinical terminology, such as SNOMED CT, ICD, or logical observation identifiers and codes (LOINC).[28] The application of conceptual modeling plus attribute binding to standard terminologies ensures that context and meaning are retained to guarantee a high degree of semantic interoperability within and between EHRs, significantly improving data quality. However, adoption of these standards by vendors has been slow due to a lack of effective regulatory or commercial mandates and incentives.

2.2. Data quality and interoperability

The value of common data models (CDM) was identified during the early 1990s. The adoption of a CDM empowered collaborative research across competing organizations.[29] This finding led to the establishment of the CDISC.[22] International collaborative research has demonstrated that semantic interoperability could be achieved by creating a CDM shared by all data contributors as these CDMs define central concepts, their attributes, constraints, and relations. CDM adoption allows for the pooling of information so that meaningful comparisons can be made.

Every EHR/EMR system is a potential data contributor and continues to make use of its own data reference model to structure its data repositories. The lack of widespread adoption of CDMs by EHR/EMR systems and issues with enforcement (governance) continues to be a major limitation. In the health-care domain, a CDM usually refers to a Clinical Data Model. openEHR has adopted a two-level modeling approach that separates its universal archetypes from applications,[30,31] as a means of optimizing semantic interoperability.

The lack of effective collaboration between ecosystem-wide stakeholders, including citizens, clients, patients, funders, health-care providers, researchers, and other institutions (data users) over time, has resulted in poor data access and data quality. Data use is limited to built-in system functionality, including reporting functionalities, as many multi-modal systems have difficulty interfacing with external systems.[32] Consequently, meaningful data aggregation to create large accessible databases or to ensure all health data pertaining to one individual is accessible through one record is limited. These represent major barriers for AI development.

As a consequence of poor quality and incomplete datasets, substantial research time, money, and effort is spent on "data cleansing" activities designed to improve data quality.[33] Data cleansing undertaken for medical AI systems can have negative effects on data quality if not performed carefully. Data cleansing can have dramatic harmful implications.[34] Stöger *et al.*[34] listed and described the following quality problems associated with the use of original data, which data cleansing activities are meant to mitigate. These are as follows:

- Absence of data – blank fields
- Dummy/default values – may be difficult to detect
- Noise (also known as the butterfly effect)
- Wrong data
- Inconsistent data
- Cryptic data
- Duplicate primary keys
- Non-unique identifiers
- Multipurpose fields
- Violation of (business) rules

Data processing sometimes requires conversion of numerical data to strings to represent a concept in words, representing another potential risk as it can lead to later issues.

2.3. Data sharing

This existing knowledge gap regarding data sharing and the need for quality data needs to be acknowledged and addressed by policy makers and research funders[35] as well as by those developing AI applications. The availability of a public library of terminology value sets enables clinical

information models and standard terminology value sets to work together to create a coherent data ecosystem.

The most accurate and adaptable method for representing computable clinical knowledge is through a dual information architecture model, which enables the development of clinical information models built from common reference components. Some existing strategies include data sharing through the use of cloud technologies and federated clinical data repositories (CDRs) to provide access to large amounts of data. CDRs need to enable reuse of data while preserving the data's original meaning and context.

Effective data sharing requires a strong data management strategy and framework including the creation of standardized, centralized processes around ingesting, classifying, storing, organizing, linking, and maintaining data. Centralization and linkage of health data on the cloud raises many security and privacy concerns as well. The use of cloud technologies to store data has the advantage of the ability to retrieve data using any type of device anytime. A major cultural shift is required to move to externally hosted services and the adoption of one set of compatible standards. CDRs need to be able to support timely health-care delivery, research, and public health initiatives as well as facilitate the creation and efficient implementation of decision-support tools. Many beneficial advances made to date are not necessarily visible to those providing frontline care.[36]

2.4. Continuing use of legacy systems

There is a desire to make the best possible use of our legacy systems to sustain existing profitable business models, to make the best possible use of significant investments made, and to maintain access to historical data. The market continues to be dominated by a few mega-EMR providers and numerous other legacy systems who are making their own data sharing arrangements, such as the HL7 Argonaut project, a private-sector initiative[37] designed to advance industry adoption of open interoperability standards. This represents a small step toward a digital transformation but is limited to users of the same enterprise-wide EMR system and its proprietary platform.

Recent collaboration managed by the Commonwealth Scientific and Industrial Research Organization's (CSIRO) Australian e-Health Research Center has resulted in the first release of the Australian Core Data for Interoperability (AUCDI) release for community comment. This collaborative consortium set out to build robust HL7 FHIR[27] profiles, extensions, and terminology value sets and bindings. This consortium's initiative (SPARKED) has launched a national FHIR Accelerator program to reinforce the move toward digital healthcare across Australia.[38] The consortium made use of universal computable clinical models (Archetypes)[39] mapped to SNOMED CT or LOINC,[40] *etc.*, which are utilized in these HL7 FHIR artifacts. The resultant AU core data set does not specify how and to what extent its elements are included in FHIR or other exchange standards. SPARKED represents another small evolutionary step toward improving data quality. While continuing to make use of legacy systems, these new initiatives need to be viewed as transitional arrangements.

3. Clinical data asset use

This review has identified a number of risk factors to be considered when extracting and collating data/information for the purpose of AI use from EHR/EMR systems. The New South Wales Government has identified these within their comprehensive AI Assurance framework[41,42] informed by groups of standards developed by the International Electrotechnical Commission (IEC)/ISO/and Joint TC (JTC1) family of SDOs. The New South Wales Government strategy includes the following key risks that need to be mitigated. These risks include:

- The use of incomplete or inaccurate data
- Having poorly defined descriptions and indicators of "fairness"
- Not ensuring ongoing monitoring of "Fairness Indicators"
- Decisions made to exclude outlier data
- Using informal or inconsistent data cleansing and repair protocols and processes
- Using informal bias detection methods
- The likelihood that re-running scenarios could produce different results (reproducibility)
- The inadvertent creation of new associations when linking data and/or metadata
- Differences between the data used for training compared to actual data
- Missing from this list was not ensuring that scenarios can be explained, which is a requirement for the generation of trustworthiness (explainability).
- Some of the questions to be answered by AI developers include:
- Is the data needed for the project in question available and of appropriate quality given the potential harms identified?
- Does your data reflect the population that will be impacted by your project or service?
- Have you considered how your AI system will address issues of diversity and inclusion (including geographic diversity)?
- Have you considered the impact regarding minority and disadvantaged groups?

- Do you have appropriate performance measures and targets?
- Do you have a way to monitor and calibrate the performance of your AI system?
- How will sensitive data be handled?

> *".…….development and implementation of AI technologies must be undertaken with appropriate consultation, transparency, accountability, and regular, ongoing review to determine its clinical and social impact and ensure it continues to benefit, and not harm, patients, health-care professionals, and the wider community."*[43]

A number of standards have been developed or are in development by the ISO/IEC/JTC1 Standards Committee 42, to assist all of us to responsibly develop, and make use of AI technologies including one for data quality for analytics and machine learning, data visualization to assess data quality and an AI data framework. The World Health Organization (WHO) has recently published its regulatory considerations on AI for health.[44]

There are numerous relevant standards for data sharing and use,[42] which cover data in general; these standards are not specific for health or clinical care data. A number of guidelines,[45-47] as well as a data governance framework,[48] have also been developed. Similar initiatives are being undertaken in other jurisdictions.[49,50] All of these measures are designed to improve data quality, streamline our use of data, and support AI development.

3.1. Data, information, and knowledge management requirements

Every known health-related terminology standard is based on an agreed categorial structure. Many of these do not identify as a formal ontology that represents a specific knowledge domain, thus resulting in ambiguities. A formal ontology consists of classes, instances, relations, functions, and axioms to reflect meaning by providing context. This allows for a clear digital understanding of concepts representing a defined knowledge domain. Terminologies were originally structured and developed to suit paper-based systems. An ontological design determines the knowledge domain's structure that enables semantic interoperability.[51]

In this digital era, it is important for standard terminologies in use to comply with the ISO standard[52] that specifies how categorial structures of terminologies need to be represented. The purpose of this ISO standard includes the need to support the development of specific standards of categorial structures for particular health-care subject fields with the minimum requirements to support meaningful exchange of information. The categorial structure approach recognizes the need for terminologies and classifications to be able to provide content related to a range of concepts and how those concepts impact the requirements of the terminology. One example is the categorial structure of nursing practice,[53] which may also be applicable to represent all types of clinical practice requiring the use of their own terminology.[54] The representation of concepts and characteristics need to especially be described in this manner for use in formal computer-based concept representation systems. Categorial structures also show relationships between categories and sub-categories.

The SNOMED CT terminology is an ontology-based comprehensive medical terminology used by many international members for standardizing the storage, retrieval, and exchange of electronic health data.[55] It is able to represent each data element and identify it together with a code. The WHO develops and updates a family of health-care terminologies including the ICD. Version 11 has an updated structure based on the use of ontology-driven tools[56] as one strategy designed to improve semantic interoperability.

Interoperability among systems requires the harmonization of such models; a project was undertaken by the Office of the National Coordinator for Health Information Technology[57] between 2017 and 2019 to advance the utility of observational data for Patient-Centered Outcomes Research (PCOR) and its interoperability across multiple networks. The PCOR project resulted in four clinical data models: (1) Sentinel, (2) PCOR Network (PCORnet), (3) Informatics for Integrating Biology and the Bedside (i2b2), and (4) Observational Medical Outcomes Partnership (OMOP).[57]

3.2. AI data quality objectives

This review has demonstrated that the design of these next-generation systems needs to be able to address the following key requirements to optimize data availability for AI development and applications:

- Ensure we have access and are able to make use of, the maximum number of data points at any required level of granularity as required to develop reliable accurate algorithms to suit AI application development. Accessing a maximum possible number of multiple desired data points needs to be achieved through linking and aggregating health-care data at scale and safely, across different tiers of care and multiple organizations, using interoperability standards and vendor-neutral data infrastructures.
- Maximize automation of routine reporting.
- Safeguard patient safety and ethical data use.

Every data point represents a single unit of information. For AI purposes, it is necessary to make use of a defined

collection of data points to determine if a pattern exists, or for algorithm development to make decisions or support decision-making or make predictions. Training any AI model requires large amounts of representative data. The number and types of accessible data points determine the accuracy of the model or a possible set of rules that can be identified. The delivery of health services is data centric where access to accurate and timely data is critical for decision-making. AI approaches making use of these data require the use of advanced analytics and access of a large amount of source data. Data-driven approaches are relevant for the provision of automated reporting as automation relies on pre-determined rules or assumptions.[58] There are significant limitations regarding access to source data collected and stored in legacy systems.

4. New and emerging technologies

This review's findings have confirmed that the interoperable and scalable ecosystem-wide architectures can be adopted, the knowledge about the health ecosystem's data supply chain, and the relationships between information models, terminologies, and ontologies with data exchange protocols. Health ecosystem-wide data supply chains need to:

(1) Include data/information flow requirements to support collaborative, person-centered life-long, and episodic continuity of care. Episodic events of multiple service episodes can also exist. Such episodes represent a treatment plan for one specific health issue such as for cancer care or a pregnancy as recorded by multiple systems over a period of time. Data collections able to meet all information needs associated with any treatment/care plan require identifiable data transfers between any number of individual and organizational health-care service providers as well as devices. Specific data needs will differ based on the individual's health status, treatment/care plans (life-long and episodic), and geographical location relative to service availability at any point in time.

(2) Facilitate the aggregation of de-identified data and identifiable data to classify any number of grouping protocols (populations) or individuals to suit specific data use cases. Data relationships will vary by use case and need to include data from systems other than data collected and stored by EHR/EMRs, such as clinical registries. Over time, such registries are expected to be generated from vendor/technology-neutral federated cloud-based health data repositories including CDRs. For some use cases, linkages may also need to include relationships between weather events or environmental status at a specific point in time or by geographical location, such as vaccination rates. CDR design needs to prioritize the separation of health information from citizen demographic or identification data by adopting a privacy-by-design approach. Every citizen needs to have control over their data and how it is used.

(3) Facilitate the linkage of health-care data with omics data, that is with the inclusion of data representing the various "omes" of an organism, to enable making sense of vast amounts of collected data to build next generations of clinical decision support and research methods and tools. At present, the use of genetic sequencing and variation information is not part of routine clinical practice because health-care professionals do not have the knowledge or skills. Most importantly, there is a lack of automated tools that can reliably associate phenotypic data from EHRs with many types of omics data to provide personal and precision care. Large-scale, well-annotated, and high-quality EHR data will have an immense impact on bringing omics and healthcare together.

(4) Facilitate the linkage of healthcare and data with the human physiome[59,60] comprising personal and mechanistic computational multi-scale models. Such models enable the provision of new types of insight into not only our understanding of human physiology and pathology but also predictions of disease and prognosis. Such insights are the result of using ontology-based EHR data linkage that parameterize these models that are able to run surprisingly reliable simulations at individual or population levels. Computational physiology and systems biology provide us with unprecedented precision to provide value-based and appropriate care as well as drive more effective drug and medical device development and faster compliance through *in silico* medicine and clinical trials.[61]

Data governance protocols, legislation, and regulations need to facilitate or enable these requirements to deliver optimal benefits of data use, including any type of effective reporting automation and AI adoption.

4.1. Next-generation EHR/EMR system characteristics

Next-generation EHR/EMR systems are designed to reduce or eliminate these gaps and improve the generation of quality data within a connected digital health ecosystem. New health platforms need to be engineered to integrate personal health information received from emerging technologies in the fields of personal health and well-being, including apps and wearables. EHR/EMR systems and CDRs should become a valuable computable data source for research and evaluation purposes as well as be enriched by data from external data sources while complying with

relevant regulatory frameworks. New systems are now adopting advanced technologies including cloud-based open (non-proprietary) ecosystem-wide platforms and openEHR's modeling approach to improve health data management. They are designed to enable plug-and-play of any number of new devices and niche applications, through architectural standards and frameworks like SMART-on-FHIR,[27] without losing the ability to share data.

Underpinned by open standards-based federated CDRs,[62] an effective separation of data and application becomes possible. The adoption of open standards enables secure access to vendor-neutral data by compliant third-party applications across the whole ecosystem. Fully standardized health data can be aggregated and utilized for many authorized purposes, including AI.

Ecosystem-wide architectural design is paramount to maximize these potential benefits. Semantic interoperability requires extensive use of ontologically structured knowledge domains[63] and ontology-driven architectures[64] as explained in details by Rector *et al.*[65] Its structure is based on the relationships between three resources: (1) Information models representing, for example, clinical concepts; (2) inference models; and (3) concept system models required to reliably undertake data abstraction – a process adopted to reduce a concept to a set of essential elements.[64,66-68]

Changing over from the legacy systems' data/information exchange paradigm to knowledge sharing at decreasing levels of abstraction requires the adoption of a reference architecture that starts at the IT concept level (semantic coordination), through the business domain concept level (agreed service function level cooperation), domain level (cross-domain cooperation), and up to individual context (skills-based end-user collaboration).[69] This architectural model supports ontology/knowledge harmonization to enable interoperability between, and integration of, systems, standards, and solutions at any level of complexity without the demand for continuous adaptation or revisions of those specifications.

Those marketing the next-generation systems have some difficulty gaining a foothold in this market as large vendors continue to protect their lucrative business models unless governments intervene. Most countries have established a national digital health framework, but these tend not to include the establishment of a suitable national supportive infrastructure designed to optimize data sharing and data quality.

4.2. Knowledge-driven architectural models and standards

The adoption of a knowledge-driven architectural model means that new-generation systems will have far greater capabilities to support life-long and person-centered care, ecosystem-wide safe data sharing through semantic interoperability, extensive automation of routine reporting, secondary data use, and advanced analytics. Widespread adoption of data standards and data governance protocols is expected to substantially reduce the need for data cleansing. Greater availability of timely, complete quality data is expected to reduce the costs of routine reporting, medical research, and other secondary data use including the development, training, and use of AI. The optimization of EHR data is expected to transform our AI capacity and the health-care industry generally.

A set of compatible standards enabling the establishment and maintenance of a well-connected national digital health ecosystem needs to be mandated. The adoption of data-driven digital health implementation strategies is now happening in some jurisdictions, such as the UK National Health Service (NHS),[70] Spain,[71] Netherlands,[72] Scandinavian countries,[73] United Arab Emirates, Kingdom of Saudi Arabia, and Jamaica,[74,75] as well as some health-care facilities. In 2019, the European Union published its Common Semantic Strategy for Health.[76]

Our work in the digital health space has identified an urgent need for new knowledge to be acquired regarding the scientific underpinnings of health data science among the health workforce.[77] Such knowledge and skills need to be applied to foster a data use culture to enable greater innovation and transformation. Only then are we in a position to improve the overall health system's performance.

This review found that essentially there are three relevant technical standards that need to be considered: openEHR,[62] HL7 FHIR,[27] and the ISO 13606:2019.[78] openEHR provides open standards for the structure, storage, and exchange of health-care information.[79] Core openEHR specifications[80] have been adopted by ISO,[78] making it a full international standard which underpins many national programs and vendor implementations worldwide.[81] The ISO 13606 standard consists of five parts and was based on the openEHR specification, making these two standards highly compatible.

4.3. openEHR

openEHR represents the evolutionary result of more than 20 years of research, innovative development, testing, implementation, and evaluation undertaken by a growing international community.

The openEHR archetypes represent health-care concepts (such as blood pressure measurement, laboratory results, and diagnoses) captured in clinical records and messages based on stable technical building blocks.[82] Its reference model defines generic but healthcare-specific

data structures, types, and value sets and a universal EHR architecture designed to process the high-level components of an EHR categorized as folder, composition, section, and entry. The entry category is further categorized into observations, evaluations, instructions, and actions. Its specifications include a platform model. Its open clinical data repository provides access to nearly 1000 archetypes to date. Collectively, these consist of the largest number of data points representing various levels of granularity in the world.[30] These archetype models are evidence-based, developed, and reviewed by well over 1000 multidisciplinary experts from 114 countries. The number of available archetypes is growing exponentially relative to openEHR-compliant implementations. The development of universal openEHR archetypes is undertaken by the potential data users who have a sound understanding of context that must be represented.

A further modeling layer is the openEHR templates which gather one or more archetypes and define use-case-specific constraints (*e.g.*, discharge summary, medication order, and clinical reports). Templates are used to drive information systems. Archetypes and templates, along with annotated clinical terminology and ontology concepts, define domain-specific information models and enable semantic interoperability in healthcare through this multi-level modeling approach. Their use for the management of clinical data especially optimizes the use of EHR data for AI purposes.

The Archetype Query Language allows formulation of portable queries using domain concepts unlike field or table names in a traditional relational database.[83] Several examples have been reported elsewhere.[84] At the knowledge level, openEHR also defines a formal clinical guideline specification (GDL) to drive decision support – all in a single standards stack.[85] There is also ongoing work to model and capture health-care processes and clinical workflows.

The openEHR Clinical Knowledge Manager (CKM) is an online clinical models repository (archetypes, templates, and clinical terminology subsets) and an advanced web-based distributed knowledge curation tool.[86] CKM supports an editorial process resembling peer-review process of a scientific journal where editors with the help of domain experts can conduct online reviews using the CKM's web interface and then publish models. Figure 1 shows the relationships between ontologies, models, and systems.

4.4. HL7 FHIR

The HL7 FHIR standard[27] represents an evolutionary result of ongoing developments of the HL7 messaging standards. Its implementation is gaining momentum. Pedrera-Jiménez *et al.*[71] explored if these three standards could work together. They found all three to be useful for the purposes for which they were designed but that they have limitations when used for different purposes. Selecting the most suitable set of standards able to best meet a defined purpose is critical.

Information models in FHIR are called resources. All clinical and other findings can only be represented using a single observation resource, which then requires adapting and bringing together many of them together as a profile to be able to represent even a simple concept like blood pressure. However, this HL7 FHIR design has been a deliberate compromise for the simplicity of technical implementation by developers at the expense of its expressivity. As a result, FHIR is being widely adopted by many vendors and health systems. There is now a common trend to use both FHIR and openEHR together, representing demographic, administrative and simple clinical data exchange using FHIR and rich clinical data using openEHR. Archetypes enable access to far more detailed clinical data points as they represent the maximal data elements for a given concept. The FHIR

Resources, on the other hand, include minimal data elements that have been adopted by current health information systems. Therefore, FHIR allows for rapid data exchange between legacy systems.

4.5. What's possible with next-generation EHR systems and data?

Current trends on building digital twins[88] exclusively using EHR data and AI without using atomic-level omics data and mechanistic knowledge and constraints of human physiology and anatomy will fall short of driving next-generation clinical decision support tools and research. Such limitations are due to not being able to train AI with reliable data. At present, available data are inevitably flawed by incorrect facts and associations as well as bias due to known shortcomings of most EHR systems in use and available data sets.

Atalag has defined an ontology mapping-based framework leveraging a multitude of existing and mature standards to bring together all these data sources (EHR, physiome, and omics) in a way that preserves the clinical, biological, physiological, and anatomical context and semantics that can drive these next-generation methods and tools,[89] as shown in Figure 2. This model represents an ontology mapping-based framework to show how compliance with various types of data exchange standards enable an openEHR-compliant clinical data repository to be populated and enable precision medicine supported by AI.

Current proprietary EMR/EHR systems and infrastructures are no longer considered "fit for purpose" due to their many shortcomings. When machine learning

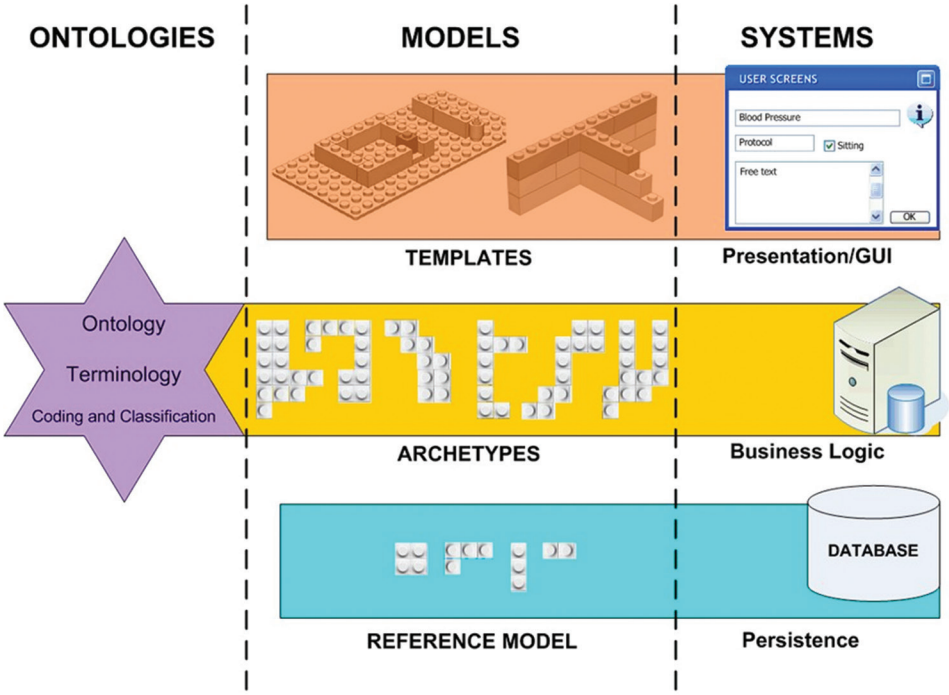


Figure 1. OpenEHR multi-level modeling[87]. Copyright © 2007 Author(s)

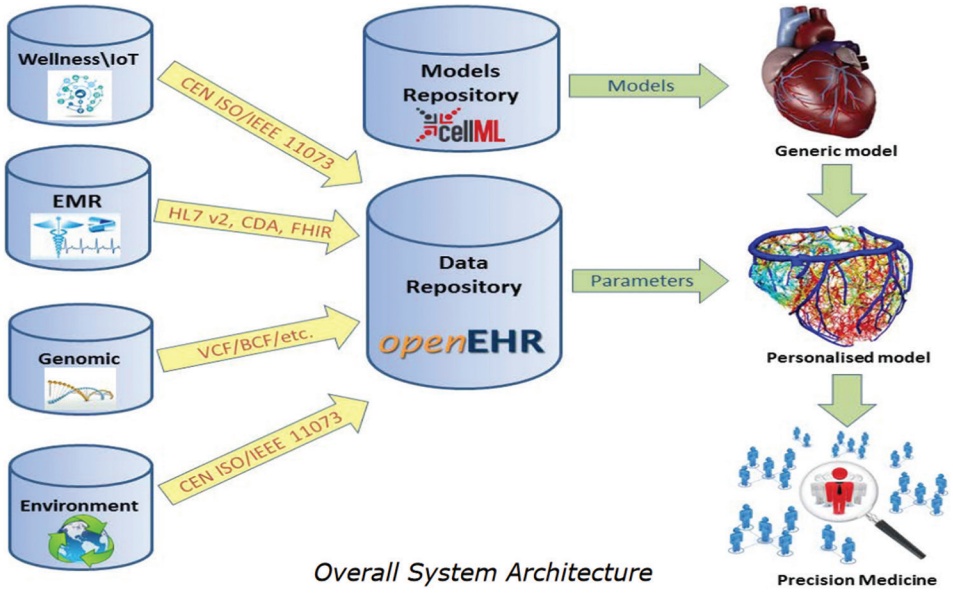


Figure 2. Ontology-based framework for an overall system architecture (adapted from a multitude of similar images developed to represent the openEHR architecture[80] and its relationships to external health data sources, exchange standards, and outcomes). Copyright © 2017 Author(s)

systems are applied to these medical data, there is an imminent and real danger of feeding AI algorithms with non-optimal EHR data which are highly likely to end up in the "garbage in garbage out" problem.[90]

There is an opportunity to capture high-quality and complete structured and computable health-care data as part of routine clinical practice. Post-ad-hoc data collection has been shown to be very expensive and error-prone, and at times, it is impossible to capture the clinical context in which the data were collected.[91] Data sources can be very diverse and range from operational EMR systems to well-structured longitudinal disease registries

Table 1. Ecosystem high-level building blocks enabling optimally connected digital health

Legislation, policy, and compliance[3]	Policy directions and associated legislation enabling the provision and compliance monitoring of health service funding mechanisms enabling universal, person-centered healthcare through an optimized primary care infrastructure supported by a well-specified and mandated digital health infrastructure. This includes privacy and cyber security legislative and regulatory requirements.
Ecosystem system architecture[20,92]	An ecosystem architecture needs to include an open platform able to meet the needs of all stakeholders and support the data supply chain. "*Better data and regular data use will create a data use culture, leading to better decisions, an improved health system, and improved health outcomes.*"[93]
Ecosystem data architecture[94]	A representation of concepts and their relationships. The data architecture defines concepts, constraints, and rules, which provide safe consistent data collection and use which retains meaning throughout the data supply chain. The domain or discourse contribute to the architectural requirements and select data from the data ecosystem based on their use case. Such data collections result in data that are structurally independent, simpler, and safer to share. The resulting lack of data silos enables advanced data analytics.
Concept representation standards[18]	Key health concepts need to be represented in the same manner throughout any digital health ecosystems to ensure data accuracy, enable consistent quality data collection at every level, optimizing data analytics, and reducing data collection burden. Consistent representation of key health concepts enables evidence-based decision-making at all levels and is best achieved by adopting a multilevel modeling approach and an open platform.
Data/information governance[48]	Specification of decision rights and an accountability framework to ensure appropriate behavior in the evaluation, creation, storage, processing, use, archiving, and deletion of information. Coordinated data governance applies to all points along the data supply chain.
Data access control	Legislation is required to indicate who can have access to identifiable and non-identifiable data for what purpose. Legislative mandates and regulatory requirements need to be considered in the light of ethical data use, and "use case" specific privacy and confidentiality, and continuity of care considerations.
Unique identifiers	An essential pre-requisite to ensure data collected can be linked to care recipients as well as to organizational and individual providers.
Cybersecurity	Minimizing risk of cyberattacks by protecting systems, servers, networks, and mobile devices. Adopt and maintain programs that educate the workforce, and manage and monitor unauthorized data access.
Vendor/technology-neutral federated data storage	The separation of systems and storage delivering scalable cost-effective data access and flexible systems for all users across the health-care network. Separating data from applications as used by the openEHR community were found to support persistent and transient data as well as real-time local and remote data access.
Electricity and broadband (Internet access) for everyone	A fundamental pre-requisite for all living in this digital era, irrespective of time, and location.

and biobanks. The patients' own contribution to their EHR, and the increasing use of mobile devices and sensors, are also important. They can add valuable insights about environmental and behavioral factors as well (*e.g.*, food, air quality, exercise, and mood).

Both data- and terminology-level standards are reasonably mature, although there is considerable overlap among certain terminology and ontology systems such as SNOMED CT and LOINC. Using fit-for-purpose data and ontology/terminology standards together can tackle most of the difficulties arising from the breadth, depth, complexity, variability, changeability, and longevity aspects of health data.

While openEHR specifications have been purposely engineered to cover all EHR data domains, including those that are intended to be exchanged by various systems, messaging standards such as HL7 v2 and HL7 FHIR have been designed to cover only data to be exchanged. These messaging standards were designed for the sake of simplicity for implementation by developers, but many of whom do not have full understanding about healthcare. Therefore, adoption of the openEHR standard is key for designing and building next-generation EHR/EMRs systems and other applications that deal with clinical data.

The emerging trend of using of HL7 FHIR beyond its purpose to represent the full breadth of clinical data in an EHR is not scalable and costly in terms of time required to develop and maintain FHIR profiles. It is far more cost-effective and safe to invest in the establishment of next-generation neutral EHR systems with vendor-neutral data repositories using openEHR and limit the use of FHIR to support simpler use cases for data exchange.

4.6. Governance and leadership

Our collective work over the last 20 plus years has highlighted the need for high-level governance leadership to maximize collaboration between all relevant stakeholders. Our collective findings over time, supported by this rapid review, have enabled us to identify the required building

blocks to make up any national foundation for successful digital health adoption. These are described in Table 1. Without such high-level government-focused ecosystem-wide collective initiatives and leadership, digital health transformation will continue to be compromised due to continuing fragmentation. The greater use of a compatible set of technology and data standards worldwide is translatable to the greater benefits and opportunities for advanced data analytics and reliable AI applications.

Every jurisdiction needs to determine how best to govern, manage, and provide strong leadership for each of these entities to facilitate the optimization of EHRs enabling AI and to meet desired health outcome objectives. Many health systems already have some of these building blocks in various forms.

5. Conclusion

This review of the current state of data availability, and data quality suitable for AI development and use, has revealed that we have a long way to go to achieve our aim of optimizing EHRs to serve as a data source for AI use. It became clear that most jurisdictions, mega-EMR vendors, and many newcomers are all tinkering at the edges by building on and working with current legacy systems and infrastructures. It is encouraging to see that some jurisdictions have bucked the trend of continuing to make incremental improvements by embarking on major digital health transformation strategies to build and implement next-generation systems and infrastructures. We have identified the need to transform high-level jurisdictional infrastructures. These infrastructural building blocks need to be designed to govern and provide strong leadership enabling ecosystem-wide compliance with mandated key sets of standards. Such standards need to enable flexibility at every point of care to ensure that data/information needs are able to be met in a timely manner for every stakeholder, and all users.

This review has identified the capacity of available next-generation technologies that need to be adopted to optimize EHR content enabling its use for AI purposes. We have the knowledge and skills required to make the best possible use of available innovative technologies to improve both the operational efficiency and effectiveness of every national health system. Many of us working in the digital health field continue to be frustrated by the lack of sufficient knowledge of the complexities associated with digital health by senior decision-makers driving investments, procurement, policy, and legislative solutions. The digital health knowledge domain is huge, both in depth and breadth. The only way we can move forward is through extensive international and multidisciplinary collaboration as made possible by the adoption of well-governed open-access standards, such as openEHR.

Multidisciplinary and multi-sector collaboration requires a change in mindset for many. The benefits of new and emerging technologies used to design and implement next-generation EHRs are huge. The use of generative AI in clinical settings can only be reliably achieved when EHRs are optimized throughout the entire global digital health ecosystem.

Acknowledgments

We thank Heather Grain, Co-Director, eHealth Education Pty Ltd for undertaking a final review and supporting this work.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization. Evelyn Hovenga
Writing – original draft Evelyn Hovenga
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. De Mello BH, Rigo SJ, Da Costa CA, *et al*. Semantic interoperability in health records standards: A systematic literature review. *Health Technol* (*Berl*). 2022;12(2):255-272.
doi: 10.1007/s12553-022-00639-w

2. Ingram D. *Health Care in the Information Society*. United Kingdom: OpenBook Publishers; 2023.
doi: 10.11647/OBP.0335

3. Hovenga E, Hullin C. Global collaborative leadership challenges and economic drivers. In: Hovenga E, Grain H, editors. *Roadmap to Successful Digital Health Ecosystems*. Ch. 3. United States: Academic Press; 2022. p. 35-63.
doi: 10.1016/B978-0-12-823413-6.00001-X

4. Heeks R. Health information systems: Failure, success and improvisation. *Int J Med Inform*. 2006;75(2):125-137.

   doi: 10.1016/j.ijmedinf.2005.07.024

5. Justinia T. The UK's National programme for IT: Why was it dismantled? *Health Serv Manage Res*. 2017;30(1):2-9.

   doi: 10.1177/0951484816662492

6. Sadoughi F, Kimiafar K, Ahmadi M, Shakeri MT. Determining of factors influencing the success and failure of hospital information system and their evaluation methods: A systematic review. *Iran Red Crescent Med J*. 2013;15(12):e11716.

   doi: 10.5812/ircmj.11716

7. Southon G, Sauer C, Dampney K. Lessons from a failed information systems initiative: Issues for complex organisations. *Int J Med Inform*. 1999;55(1):33-46.

   doi: 10.1016/s1386-5056(99)00018-0

8. Johnstone R. *Global Report Reveals Senior Officials Lack Understanding to Drive Digital Transformation of Government*. Singapore: Global Government Forum; 2022 Available from: https://www.globalgovernmentforum.com/global-report-reveals-senior-officials-lack-understanding-to-drive-digital-transformation-of-government [Last accessed on 2024 Mar 09].

9. Wallace M, Sharfstein JM. The patchwork U.S. Public health system. *N Engl J Med*. 2022;386(1):1-4.

   doi: 10.1056/NEJMp2104881

10. Steinkamp J, Kantrowitz JJ, Airan-Javia S. Prevalence and sources of duplicate information in the electronic medical record. *JAMA Netw Open*. 2022;5(9):e2233348.

    doi: 10.1001/jamanetworkopen.2022.33348

11. Wang EC, Wright A. Characterizing outpatient problem list completeness and duplications in the electronic health record. *J Am Med Inform Assoc*. 2020;27(8):1190-1197.

    doi: 10.1093/jamia/ocaa125

12. Madden JM, Lakoma MD, Rusinak D, Lu CY, Soumerai SB. Missing clinical and behavioral health data in a large electronic health record (EHR) system. *J Am Med Inform Assoc*. 2016;23(6):1143-1149.

    doi: 10.1093/jamia/ocw021

13. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*. 2013;46(5):830-836.

    doi: 10.1016/j.jbi.2013.06.010

14. Downing NL, Bates DW, Longhurst CA. Physician burnout in the electronic health record era: Are we ignoring the real cause? *Ann Intern Med*. 2018;169(1):50-51.

    doi: 10.7326/M18-0139

15. Zhang J, Symons J, Agapow P, *et al*. Best practices in the real-world data life cycle. *PLoS Digit Health*. 2022;1(1):e0000003.

    doi: 10.1371/journal.pdig.0000003

16. Zhang J, Morley J, Gallifant J, *et al*. Mapping and evaluating national data flows: Transparency, privacy, and guiding infrastructural transformation. *Lancet Digit Health*. 2023;5(10):e737-e748.

    doi: 10.1016/S2589-7500(23)00157-7

17. HL7. *Health Level Seven International*. Available from: https://www.hl7.org [Last accessed on 2019 Feb 06].

18. SNOMED-International. *Leading Healthcare Terminology, Worldwide*. Available from: https://www.snomed.org [Last accessed on 2024 Mar 10].

19. DICOM. *Digital Imaging and Communications in Medicine*. Available from: https://www.dicomstandard.org [Last accessed on 2019 May 23].

20. Beale T, Hovenga E. The knowledge-driven platform: Strategic technologies for a platform ecosystem approach. In: Hovenga E, Grain H, editors. *Roadmap to Successful Digital Health Ecosystems*. Ch. 6. United States: Academic Press; 2022. p. 115-141.

    doi: 10.1016/B978-0-12-823413-6.00006-9

21. Hovenga E, Grain H, Beale T. Fragmented global standards development organisations. In: Hovenga E, Grain H, editors. *Roadmap to Successful Digital Health Ecosystems*. Ch. 4. United States: Academic Press; 2022. p. 65-96.

    doi: 10.1016/B978-0-12-823413-6.00025-2

22. CDISC. *Clinical Data Interchange Standards Consortium*. Available from: https://www.cdisc.org [Last accessed on 2019 May 23].

23. JIC. *Joint Initiative Council for Global Health Informatics Standardisation*. Available from: https://www.jointinitiativecouncil.org [Last accessed on 2024 Mar 10].

24. ISO/TS_21564. *2019 Health Informatics-Terminology Resource Map Quality Measures (MapQual)*. Switzerland: Intenational Organisation for Standardisation; 2019. Available from: https://www.iso.org/standard/71088.html [Last accessed on 2019 Dec 06].

25. OpenEHR. *Archetype Technology Overview*. London: OpenEHR Foundation. Available from: https://specifications.openehr.org/releases/am/latest/overview.html [Last accessed on 2021 Jun 11].

26. HL7-International. *HL7 Version 3 Clinical Document Architecture (CDA) Release 2*. Available from: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7 [Last accessed on 2018 Sep 26].

27. HL7-International. *Fast Healthcare Interoperability Resources Specification (FHIR)*. Available from: https://www.hl7.org/fhir/overview.html [Last accessed on 2024 Feb 28].

28. Atalag K, Yang HY, Tempero E, Warren J. Model driven development of clinical information sytems using openEHR.

Stud Health Technol Inform. 2011;169:849-853.

29. Weeks J, Pardee R. Learning to share health care data: A brief timeline of influential common data models and distributed health data networks in U.S. health care research. *EGEMS* (*Wash DC*). 2019;7(1):4.

doi: 10.5334/egems.279

30. OpenEHR/CKM. *Clinical Knowledge Manager*. Available from: https://ckm.openehr.org/ckm/# [Last accessed on 2019 Feb 21].

31. OpenEHR-Foundation. *OpenEHR Reference Model (RM)-Latest*. London: OpenEHR Foundation; 2017. Avaialble from: https://openehr.org [Last accessed on 2024 Jul 23].

32. Cresswell KM, Mozaffar H, Lee L, Williams R, Sheikh A. Safety risks associated with the lack of integration and interfacing of hospital health information technologies: A qualitative study of hospital electronic prescribing systems in England. *BMJ Qual Saf*. 2017;26(7):530-541.

doi: 10.1136/bmjqs-2015-004925

33. Guo M, Wang Y, Yang Q, *et al*. Normal workflow and key strategies for data cleaning toward real-world data: Viewpoint. *Interact J Med Res*. 2023;12:e44310.

doi: 10.2196/44310

34. Stöger K, Schneeberger D, Kieseberg P, Holzinger A. Legal aspects of data cleansing in medical AI. *Comput Law Secur Rev*. 2021;42:105587.

doi: 10.1016/j.clsr.2021.105587

35. Scott P, Dunscombe R, Evans D, Mukherjee M, Wyatt J. Learning health systems need to bridge the "two cultures" of clinical informatics and data science. *J Innov Health Inform*. 2018;25:126-131.

doi: 10.14236/jhi.v25i2.1062

36. Cresswell K, Domínguez Hernández A, Williams R, Sheikh A. Key challenges and opportunities for cloud technology in health care: Semistructured interview study. *JMIR Hum Factors*. 2022;9(1):e31246.

doi: 10.2196/31246

37. HL7. *Argonaut Project 2019*. Available from: http://docs.smarthealthit.org/argonaut [Last accessed on 2019 Feb 06].

38. CSIRO-eHealth-Research-Centre. *SPARKED-AU FHIR Accelerator Brisbane*; 2024. Available from: https://confluence.csiro.au/display/fhir/sparked+-+au+fhir+accelerator [Last accessed on 2024 Feb 17].

39. ISO-13606-2. *Health Informatics--Electronic Health Record Communication--Part 2: Archetype Interchange Specification: International Organsiation for Standardisation*; 2008. Available from: https://www.iso.org/standard/50119.html [Last accessed on 2019 May 25].

40. LOINC. *Logical Observation Identifiers Names and Codes from Regenstrief*. Indiana: Regenstrief. Available from: https://loinc.org [Last accessed on 2019 Feb 25].

41. NSW-Government DN. AI Guidelines: Using Artifical Intelligence in the NSW Public Sector; 2023. Available from: https://www.education.gov.au/schooling/announcements/australian-framework-generative-artificial-intelligence-ai-schools [Last accessed on 2024 Feb 16].

42. NSW-Government. *AI Assurance Framework;* 2022. Available from: https://www.digital.nsw.gov.au/policy/artificial-intelligence/nsw-artificial-intelligence-assurance-framework [Last accessed on 2024 Feb 16].

43. AMA. *Position Statement: Artifical Intelligence in Healthcare*; 2023. Available from: https://www.ama.com.au/sites/default/files/2023-08/artificial%20intelligence%20in%20healthcare%20-%20ama.pdf [Last accessed on 2024 Feb 16].

44. World Health Organisation. *Regulatory Considerations on Artificial Intelligence for health*. Geneva: World Health Organization; 2023. Available from: https://iris.who.int/handle/10665/373421 [Last accessed on 2024 Feb 16].

45. Australian-Government DoI, Science and Resources. *Australia's AI Ethics Principles*. Available from: https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles [Last accessed on 2024 Feb 16].

46. Australian-Government SDA. *Guidelines for Secure AI System Development*; 2023. Available from: https://www.cyber.gov.au/about-us/view-all-content/advice-and-guidance/guidelines-secure-ai-system-development [Last accessed on 2024 Feb 17].

47. Australian-Government-DoH. *National Health Reform Agreement (NHRA) 2020-25*. Available from: https://www.health.gov.au/initiatives-and-programs/2020-25-national-health-reform-agreement-nhra [Last accessed on 2024 Feb 17].

48. Data-Governance-Institute. *The DGI-Data Governance Framework*. Available from: https://datagovernance.com [Last accessed on 2024 Feb 16].

49. The-White-House. *Delivering on the Promise of AI to Improve Health Outcomes*; 2023. Available from: https://www.whitehouse.gov/briefing-room/blog/2023/12/14/delivering-on-the-promise-of-ai-to-improve-health-outcomes [Last accessed on 2024 Feb 28].

50. Government-of-the-Netherlands. *Dutch Government Presents Vision on Generative AI*; 2024. Available from: https://www.government.nl/latest/news/2024/01/18/dutch-government-dutch-government-presents-vision-on-generative-ai#:~:text=the%20dutch%20government%20has%20announced,via%20the%20dutch%20ai%20coalition [Last accessed on 2024 Feb 28].

51. Elkin P, editor. *Terminology, Ontology and their Implementations*. 2nd ed. Berlin: Springer; 2023.

52. ISO17115. *Health Informatics-Representation of Categorial Structures of Terminology* (*CatStructure*); 2020. Available

from: https://www.iso.org/standard/73695.html [Last accessed on 2023 Dec 05].

53. ISO18104. *Health Informatics-Categorial Structures for Representation of Nursing Practice in Terminological Systems*. Switzerland: International Organisation of Standards; 2023. Available from: https://www.iso.org/standard/81132.html [Last accessed on 2023 Dec 05].

54. Hovenga E. Integrating a categorial structure for clinical practice into EHRs. *Stud Health Technol Inform*. 2024;310:74-78.

doi: 10.3233/SHTI230930

55. El-Sappagh S, Franda F, Ali F, Kwak KS. SNOMED CT standard ontology based on the ontology for general medical science. *BMC Med Inform Decis Mak*. 2018;18(1):76.

doi: 10.1186/s12911-018-0651-5

56. Rodrigues JM, Kumar A, Bousquet C, Trombert B. Using the CEN/ISO standard for categorial structure to harmonise the development of WHO international terminologies. *Stud Health Technol Inform*. 2009;150:255-259.

57. ONC. *Common Data Model Harmonisation*. Available from: https://www.healthit.gov/topic/scientific-initiatives/pcor/common-data-model-harmonization-cdm [Last accessed on 2024 Feb 21].

58. Aldoseri A, Al-Khalifa KN, Hamouda AM. Re-thinking data strategy and integration for artificial intelligence: Concepts, opportunities, and challenges. *Appl Sci*. 2023;13(12):7082.

doi: 10.3390/app13127082

59. McCormick KA. The digital needs of genomics resulting from pandemics. In: Hovenga E, Grain H, editors. *Roadmap to Successful Digital Health Ecosystems*. Ch. 15. United States: Academic Press; 2022. p. 363-374.

doi: 10.1016/B978-0-12-823413-6.00018-5

60. Hassan M, Awan FM, Naz A, *et al*. Innovations in genomics and big data analytics for personalized medicine and health care: A review. *Int J Mol Sci*. 2022;23(9):4645.

doi: 10.3390/ijms23094645

61. Viceconti M, Hunter P, Hose R. Big data, big knowledge: Big data for personalized healthcare. *IEEE J Biomed Health Inform*. 2015;19(4):1209-1215.

doi: 10.1109/JBHI.2015.2406883

62. OpenEHR. *An Open Domain-Driven Platform for Developing Flexible e-Health systems*. London: OpenEHR Foundation. Available from: https://www.openehr.org [Last accessed on 2019 2019 Feb 06].

63. Blobel B, Kalra D, Koehn M, *et al*. The role of ontologies for sustainable, semantically interoperable and trustworthy EHR solutions. *Stud Health Technol Inform*. 2009;150:953-957.

64. Rector AL, Rogers J, Taweel A. Models and inference methods for clinical systems: A principled approach. *Stud Health Technol Inform*. 2004;107(Pt 1):79-83.

65. Rector A, Schulz S, Rodrigues JM, Chute CG, Solbrig H. On beyond gruber: "Ontologies" in today's biomedical information systems and the limits of OWL. *J Biomed Inform*. 2019;100:100002.

doi: 10.1016/j.yjbinx.2019.100002

66. Rector A, Johnson P, Tu SW, Wroe C, Rogers J, editors. *Interface of Inference Models with Concept and Medical Record Models. In: AIME'01: Proceedings of the 8th Conference on AI in Medicine in Europe: Artifical Intelligence Medicine*; 2001.

doi: 10.1007/3-540-48229-6_43

67. Rector AL. The interface between information, terminology, and inference models. *Stud Health Technol Inform*. 2001;84(Pt 1):246-250.

68. Rector A, Marley T, Qamar R. *Models of use and Model of Meaning Slideplayer*; 2017. Available from: https://slideplayer.com/slide/11210854 [Last accessed on 2024 Feb 29].

69. ISO-23903. *Health Informatics-Interoperability and Integration Reference Architecture-Model and Framework*. Geneva: International Organization for Standards (ISO); 2021. Available from: https://www.iso.org/standard/77337.html [Last accessed on 2024 Jun 11].

70. NHS-England. *Digital Transformation*. Available from: https://www.england.nhs.uk/digitaltechnology [Last accessed on 2024 Feb 28].

71. Pedrera-Jiménez M, García-Barrio N, Frid S, *et al*. Can OpenEHR, ISO 13606, and HL7 FHIR work together? An agnostic approach for the selection and application of electronic health record standards to the next-generation health data spaces. *J Med Internet Res*. 2023;25:e48702.

doi: 10.2196/48702

72. OECD. *Towards an Integrated Health Information System in the Netherlands*. Paris: OECD Publishing; 2022.

doi: 10.1787/a1568975-en

73. Eit-Health E. *Implementing the European Health Data Space in Sweden*; 2023. Available from: https://eithealth.eu/wp-content/uploads/2023/10/implementing-the-european-health-data-space-in-sweden.pdf [Last accessed on 2024 Feb 28].

74. Government-of-Jamaica. *Jamaica Poursues Digital Transformation of Health Sector: Ministry of Health and Welness*; 2024 Available from: https://www.moh.gov.jm/jamaica-pursues-digital-transformation-of-health-sector [Last accessed on 2024 Feb 21].

75. Jones M. Caribbean/PAHO-Jamaican case study. In: Hovenga E, Grain H, editors. *Roadmap to Successful Digital Health Ecosystems*. Ch. 4. United States: Academic Press; 2022. p. 523-535.

doi: 10.1016/B978-0-12-823413-6.00026-4

76. eHAction. *Common Semantic Strategy for Health in the European Union*; 2019 Available from: http://ehaction.eu/eu-common-semantic-strategy-open-consultation [Last accessed on 2024 Feb 28].

77. CODE-Center-for-Open-Data-Enterprise. *Sharing and Utilizing Health Data for AI Applications-Round Table Report*; 2019. Available from: https://www.hhs.gov/sites/default/files/sharing-and-utilizing-health-data-for-ai-applications.pdf [Last accessed on 2024 Feb 22].

78. ISO13606. *The ISO 13606 Standard Explained*; 2019. Available from: http://www.en13606.org/information.html [Last accessed on 2024 Feb 18].

79. OpenEHR. *Specifications*. Available from: https://specifications.openehr.org [Last accessed on 2021 Apr 25].

80. Atalag K, Beale T, Chen R, Gornik T, Heard S, McNicoll I. *OpenEHR-A Semantically-Enabled, Vendor-Independent Health Computing Platform-White Paper*; 2017. Avaialble from: https://www.openehr.org/resources/white_paper_docs/openEHR_vendor_independent_platform.pdf [Last accessed on 2017 Feb 17].

81. ISO-13606-1. *Health Informatics--Electronic Health Record Communication--Part 1: Reference Model*. Switzerland: ISO; 2019. Available from https://www.iso.org/standard/67868.html [Last accessed on 2017 Feb 18].

82. Beale T. Archetypes: Constraint-based domain models for future-proof information systems. In: Baclawski K, Kilov H, editors. *Eleventh OOPSLA Workshop on Behavioral Semantics: Serving the Customer*. Boston: Northeastern University; 2002.

83. Ma C, Frankel H, Beale T, Heard S. EHR query language (EQL)--a query language for archetype-based health records. *Stud Health Technol Inform*. 2007;129(Pt 1):397-401.

84. OpenEHR. *Archetype Query Language (AQL)*. Available from: https://specifications.openehr.org/releases/query/latest/aql.html [Last accessed on 2019 Feb 12].

85. Chen R, Valladares C, Corbal I, Anani N, Koch S. Early experiences from a guideline-based computerized clinical decision support for stroke prevention in atrial fibrillation. *Stud Health Technol Inform*. 2013;192:244-247.

86. Garde S. Clinical knowledge governance: The international perspective. *Stud Health Technol Inform*. 2013;193:269-281.

87. Atalag K, Bilgen S. *Multi-level Modeling and the Role of Archetypes in the Design of Health Information Systems: A Modeling Example in Endoscopy. In: Conference proceedings of the 2007 International Symposium on Health Informatics and Bioinformatics HIBIT07, Atalya, Turkey*.

doi: 10.13140/2.1.4762.2727

88. Viceconti M, De Vos M, Mellone S, Geris L. Position paper from the digital twins in healthcare to the virtual human twin: A moon-shot project for digital health research. *IEEE J Biomed Health Inform*. 2023;28:491-501.

doi: 10.1109/JBHI.2023.3323688

89. Nickerson D, Atalag K, De Bono B, *et al*. The human physiome: How standards, software and innovative service infrastructures are providing the building blocks to make it achievable. *Interface Focus*. 2016;6(2):20150103.

doi: 10.1098/rsfs.2015.0103

90. Teno JM. Garbage in, Garbage out-words of caution on big data and machine learning in medical practice. *JAMA Health Forum*. 2023;4(2):e230397.

doi: 10.1001/jamahealthforum.2023.0397

91. Institute of Medicine Committee on Quality of Health Care in America. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: National Academies Press.

92. ISO/TR14639. *Briefing Note: Strengthening National Health Systems through a Capacity-Based eHealth Architecture*. Switzerland: International Organization of Standards ISO; 2014. Available from: https://www.iso.org/files/live/sites/isoorg/files/archive/pdf/en/14639-brochureversionv7.pdf [Last accessed on 2020 Jul 18].

93. PATH. *Data Use partnerships: Theory of Change*; 2016. Available from: https://www.path.org/our-impact/resources/data-use-partnership-theory-of-change [Last accessed on 2021 Jul 22].

94. Castro A, Machado J, Roggendorf M, Soller H. *How to build a data architecture to Drive Innovation-Today and Tomorrow*. *McKinsey Technology*; 2021. Available from: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/how-to-build-a-data-architecture-to-drive-innovation-today-and-tomorrow [Last accessed on 2021 Apr 12].

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

PERSPECTIVE ARTICLE

Artificial intelligence for ophthalmic drug discovery and development: Capabilities, applications, and challenges

Siddharth Gandhi[1] and Michael Balas[2]*

[1]School of Medicine, Queen's University, Kingston, Ontario, Canada

[2]Department of Ophthalmology & Vision Sciences, University of Toronto, Toronto, Ontario, Canada

Abstract

The integration of artificial intelligence (AI) into ophthalmic drug discovery and development presents transformative opportunities to address the inherent complexities and challenges of creating targeted therapies for eye diseases. The ability of AI to process vast datasets can facilitate the discovery of novel drug candidates, improve predictions of drug efficacy and safety, and streamline the drug development pipeline. Applications can range from enhancing target identification and compound screening to refining predictive toxicology. However, challenges such as data limitations, computational demands, model interpretability, and ethical considerations remain. Despite these hurdles, the integration of AI with emerging technologies and its potential to optimize clinical trials signifies a new era of innovation in ophthalmology, emphasizing its critical role in addressing current challenges and advancing therapeutic development. In this paper, we explore the role of AI in ophthalmic drug discovery, highlighting its potential to address critical challenges in the field and delineating its impact across various stages of drug development.

Keywords: Ophthalmology; Artificial intelligence; Drug discovery; Drug development

***Corresponding author:**
Michael Balas
(michael.balas@mail.utoronto.ca)

Citation: Gandhi S, Balas M. Artificial intelligence for ophthalmic drug discovery and development: Capabilities, applications, and challenges. *Artif Intell Health*. 2024;1(3):26-30.
doi: 10.36922/aih.3341

Received: April 1, 2024

Accepted: May 13, 2024

Published Online: July 22, 2024

1. Introduction

The field of ophthalmology encounters unique challenges in drug development stemming from the complexity of eye diseases and the necessity for highly targeted therapeutic interventions. Traditional drug discovery processes are often slow and costly, and they are hampered by high failure rates, particularly in predicting drug efficacy and safety within the context of eye diseases.[1] These challenges underscore the urgent need for innovative approaches to accelerate and refine drug discovery in ophthalmology.

Artificial intelligence (AI) is reshaping the frontier of drug discovery. By analyzing large datasets, AI can efficiently identify patterns and insights that may lead to the discovery of novel drug candidates, as well as enhance predictions of drug efficacy and toxicities. Moreover, AI can enhance our understanding of disease mechanisms, contribute to personalized medicine, and streamline drug development pipelines, thereby expediting the process, reducing costs, and ultimately increasing effectiveness.[2]

The integration of AI into ophthalmic drug discovery represents a convergence of technology and medicine that could significantly advance the treatment of eye diseases. By leveraging AI in target identification, compound screening, and predictive toxicology, researchers can overcome some of the traditional bottlenecks in the drug discovery pipeline, facilitating much-needed treatments for patients with eye diseases.[1] Beyond expediting the development of safe, effective, and targeted treatments, this integration may also enrich our understanding of ophthalmic disease processes. In this paper, we explore the role of AI in ophthalmic drug discovery, offering insights into its potential to address the critical challenges in the field.

2. AI in drug discovery

The discovery and validation of therapeutic targets are critical steps in the drug discovery pipeline. This process involves identifying molecular targets, such as proteins that play a key role in disease pathogenesis, and confirming their suitability for therapeutic intervention.[3] Traditional methods for target identification are often labor-intensive, time-consuming, and fraught with uncertainty. AI can analyze vast datasets from genomic, proteomic, and other multi-omic studies to uncover potential targets related to eye diseases more efficiently than traditional approaches.[3] Once potential targets are identified, AI can further assist in validating these targets by predicting their role in disease progression and response to therapeutic intervention, thus enhancing the specificity and effectiveness of drug development efforts in ophthalmology.[3]

After identifying possible therapeutic targets, the process of compound screening involves testing numerous agents for activity against the identified target. This phase also includes optimizing these agents to improve their efficacy, safety profile, and pharmacokinetic properties.[2] AI has the potential to significantly accelerate this phase using models that predict compound-target interactions, thereby narrowing down the vast library of potential compounds to those most likely to exhibit desired therapeutic effects. In addition, AI-driven models can simulate the molecular docking process and provide predictions on how different compounds will bind to target proteins.[4] This not only accelerates the screening process but also enhances the precision of compound selection for further development. Moreover, AI can optimize compound structures by predicting modifications that enhance drug-like properties, ensuring that the most promising candidates are advanced to the next stages of drug development with optimized profiles for effectiveness and safety.[2]

Finally, predictive toxicology is essential in assessing the safety profile of drug candidates before they advance to clinical trials, given the substantial costs associated with late-stage drug development failures due to unforeseen toxicity.[5] AI models, especially those trained on extensive databases of chemical structures and their associated toxicological profiles, offer high accuracy in predicting the toxicity of new compounds.[5] By integrating data from various sources, including preclinical studies and known drug safety profiles, AI can forecast adverse effects and serve as an early warning system to prioritize compounds with favorable safety profiles.[6] This capability not only reduces the risk of late-stage failures but also ensures a more efficient allocation of resources toward candidates with the highest likelihood of success in treating ophthalmic conditions. It is important to note that while specific applications of AI models in ophthalmology are still emerging, the theoretical and operational frameworks established in other therapeutic areas provide a promising foundation for their application in ocular drug development.

3. Current applications and case studies in ophthalmology

As patients' adherence to dosing regimens (eye drops) and frequent intraocular injections can be substantial barriers to effective chronic ocular disease management, sustained drug delivery strategies can be helpful.[7] However, as these sustained therapeutic methods are traditionally achieved by implantable devices, there is a risk of excipient material buildup, the need for device removal, potential adverse reactions, etc.[7] An alternative approach would be to increase the retention time and therapeutic effects of drugs in the eye without using implants, as attempted by Hsueh *et al.*[7] As ocular melanin has a low turnover rate, they hypothesized that a melanin-binding peptide could be conjugated to small-molecule drugs to increase their retention time and therapeutic effect. Since incorporating multiple functions into a single peptide sequence is challenging, they used machine learning methods to help engineer peptide sequences that could simultaneously provide these desired functions. As a result, their engineered peptide exhibited increased cell-penetrating properties and high melanin binding capacity while demonstrating low cytotoxicity. They tested these compounds in rabbits and discovered that their multifunctional peptide greatly enhanced the pharmacokinetics and pharmacodynamics of brimonidine when compared to normal use. In this work, machine learning played a key role in identifying important variables for desired peptide function, refining peptide design, and achieving desired therapeutic goals.

The application of AI extends beyond enhancing drug delivery systems to revolutionizing our approach to disease management strategies, including neurodegenerative

ocular diseases. There is a plethora of neurodegenerative conditions that can cause damage to the optic nerve.[8] One of the primary pathophysiological mechanisms of action involves the damage of retinal ganglion cell (RGC) axons.[8] There is emerging evidence suggesting that axonal damage can initiate RGC death through reactive oxygen species (ROS), which in turn increases disulfide bond formation between cysteine side chains to cause further cellular damage.[8] Redox-active phosphine-borane complexes have been proposed as protective molecules that can activate cellular pathways to prevent these disulfide bonds from forming.[8] However, limited pharmacological data exists for these compounds. To resolve this issue, Remtulla *et al.*[8] trained neural networks on features such as cellular permeability, oral absorption, blood–brain barrier permeability, and serum protein binding to reliably predict the pharmacokinetics of boron-containing compounds. Their results revealed that phosphine-boron compounds met the necessary pharmacokinetic profile to function as orally active drug candidates. Ultimately, this study underscores the innovative use of machine learning in evaluating the pharmacokinetics of emerging compounds, such as phosphine-borane complexes, advancing their potential as neuroprotective agents against RGC damage. It exemplifies the ability to generate new perspectives in ocular pharmacology using pre-existing data and AI algorithms.

4. Challenges and limitations

Despite the promising advancements and successful applications of AI in ophthalmologic drug discovery, several challenges and limitations remain that require acknowledgment and resolution. First and foremost, the quality and quantity of data available for AI models significantly influence their performance and reliability.[9] In the realm of ophthalmology, high-quality, diverse, and annotated datasets, especially from clinical settings, are often scarce or fragmented.[9] This limitation can lead to biases in AI models, reducing their generalizability and accuracy when applied to broader, more diverse populations.

Furthermore, the computational resources required for AI research are substantial. The processing of large datasets and the training of sophisticated models necessitate advanced hardware and significant computational power, which can be a barrier for institutions with limited resources.[10] This technological and financial barrier may lead to disparities in research advancements and the adoption of AI technologies across different regions and institutions.

Another significant challenge is the interpretability of AI models, particularly those based on deep learning algorithms. These models are often described as "black boxes" due to their opaque decision-making processes, which are difficult for humans to comprehend.[11] This lack of transparency can hinder the trust and acceptance of AI-driven discoveries among clinicians, researchers, and regulatory bodies, which is critical for translating AI discoveries into practical therapeutic interventions. To address these challenges, the current strategies focus on the development of explainable AI methods, such as feature importance scores and rule-based decision trees, designed to demystify AI decisions and enhance model transparency. In addition, integrating domain-specific knowledge and employing hybrid models that combine deep learning with interpretable statistical methods are proving crucial in improving both the interpretability and reliability of these systems. Ensuring robustness and generalization through rigorous testing, coupled with proactive stakeholder engagement, is essential to validating and gaining acceptance for AI technologies in clinical settings.

The integration of AI into drug discovery also presents ethical and regulatory challenges. The use of patients' data raises privacy concerns, requiring stringent data protection measures and ethical oversight to ensure patient confidentiality and consent.[3] Moreover, regulatory frameworks for AI-assisted drug discovery and development are still in their infancy, lacking clear guidelines for validation, approval, and oversight of AI-driven methodologies. This regulatory uncertainty can delay the adoption and application of AI technologies in ophthalmology drug discovery.

5. Future direction

The integration of AI into ophthalmologic drug discovery marks a new era of medical innovation and operational efficiency, addressing longstanding challenges and opening new avenues for therapeutic development. Looking ahead, several research objectives are set to further leverage the capabilities of AI systems. Among these, a key goal will involve employing these technologies to enhance our understanding of complex eye diseases at the molecular level. Future efforts are likely to focus on developing more sophisticated algorithms that can process and analyze the increasingly large and complex datasets generated by biomedical research.[9] This will not only improve the accuracy of target identification and validation but also enable the discovery of novel biomarkers and therapeutic targets.[2]

Another promising direction involves the integration of AI with other emerging technologies, such as gene editing and stem cell therapy.[12] By combining AI's predictive

and analytical capabilities with these novel therapeutic approaches, researchers can accelerate the development of personalized medicine strategies for ophthalmic diseases.

Moreover, AI is set to play a crucial role in overcoming the challenges associated with the clinical trial phase of drug development.[13] By predicting patient responses to potential treatments and identifying the most suitable candidates for participation in trials, AI can streamline the process of participant recruitment. This would, in turn, lead to faster, more cost-effective trials that also enable researchers to fine-tune dosage and treatment protocols earlier, reducing the probability of unforeseen adverse reactions and late-stage trial failures.[13]

6. Conclusion

The integration of AI into ophthalmic drug discovery and development represents a paradigm shift, offering novel approaches to longstanding challenges in this field. Using AI in target identification, compound screening, and predictive toxicology, we are on the brink of making substantial progress in the treatment of eye diseases. The successful applications and case studies discussed underscore AI's potential to enhance drug delivery systems, refine disease management, and expedite drug development. However, addressing challenges and limitations, including data quality, computational resource requirements, model interpretability, and regulatory issues, is paramount to fully realizing its transformative potential. Nonetheless, the continued evolution of AI technologies, coupled with their integration into emerging therapeutic modalities, is likely to pave the way for advancements that will redefine our approach to treating ophthalmic diseases, ultimately improving patient outcomes and quality of life.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Michael Balas
Writing – original draft: Siddharth Gandhi
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. Wang N, Zhang Y, Wang W, *et al*. How can machine learning and multiscale modeling benefit ocular drug development? *Adv Drug Deliv Rev*. 2023;196:114772.

   doi: 10.1016/j.addr.2023.114772

2. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. *Drug Discov Today*. 2021;26(1):80-93.

   doi: 10.1016/j.drudis.2020.10.010

3. Pun FW, Ozerov IV, Zhavoronkov A. AI-powered therapeutic target discovery. *Trends Pharmacol Sci*. 2023;44(9):561-572.

   doi: 10.1016/j.tips.2023.06.010

4. Crampon K, Giorkallos A, Deldossi M, Baud S, Steffenel LA. Machine-learning methods for ligand-protein molecular docking. *Drug Discov Today*. 2022;27(1):151-164.

   doi: 10.1016/j.drudis.2021.09.007

5. Wang MWH, Goodman JM, Allen TEH. Machine learning in predictive toxicology: Recent applications and future directions for classification models. *Chem Res Toxicol*. 2021;34(2):217-239.

   doi: 10.1021/acs.chemrestox.0c00316

6. Yang S, Kar S. Application of artificial intelligence and machine learning in early detection of adverse drug reactions (ADRs) and drug-induced toxicity. *Artif Intell Chem*. 2023;1(2):100011.

   doi: 10.1016/j.aichem.2023.100011

7. Hsueh HT, Chou RT, Rai U, *et al*. Machine learning-driven multifunctional peptide engineering for sustained ocular drug delivery. *Nat Commun*. 2023;14(1):2509.

   doi: 10.1038/s41467-023-38056-w

8. Remtulla R, Das SK, Levin LA. Predicting absorption-distribution properties of neuroprotective phosphine-borane compounds using *in silico* modeling and machine learning. *Molecules*. 2021;26(9):2505.

   doi: 10.3390/molecules26092505

9. Li Z, Wang L, Wu X, *et al*. Artificial intelligence in ophthalmology: The path to the real-world clinic. *Cell Rep Med*. 2023;4(7):101095.

   doi: 10.1016/j.xcrm.2023.101095

10. Jia Z, Chen J, Xu X, *et al*. The importance of resource awareness in artificial intelligence for healthcare. *Nat Mach*

Intell. 2023;5(7):687-698.

doi: 10.1038/s42256-023-00670-0

11. Balagurunathan Y, Mitchell R, El Naqa I. Requirements and reliability of AI in the medical context. *Phys Med*. 2021;83:72-78.

doi: 10.1016/j.ejmp.2021.02.024

12. Nosrati H, Nosrati M. Artificial intelligence in regenerative medicine: Applications and implications. *Biomimetics (Basel)*. 2023;8(5):442.

doi: 10.3390/biomimetics8050442

13. Zhang B, Zhang L, Chen Q, Jin Z, Liu S, Zhang S. Harnessing artificial intelligence to improve clinical trial design. *Commun Med* (*Lond*). 2023;3(1):191.

doi: 10.1038/s43856-023-00425-3

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Predicting mortality outcomes in individual COVID-19 patients using machine learning algorithms

Nikolaos Kourmpanis*, Joseph Liaskos, Emmanouil Zoulias, and John Mantas

Laboratory of Health Informatics, Department of Public Health, Faculty of Nursing, National and Kapodistrian University of Athens, Athens, Greece

Abstract

In late 2019, the COVID-19 disease emerged, caused by the SARS-CoV-2 virus, and has since spread worldwide, becoming a global pandemic and resulting in almost seven million deaths to date. In addressing this global crisis, artificial intelligence has played a crucial role, particularly through the development of predictive models using machine learning algorithms, which have been successfully applied to solving a multitude of problems across multiple scientific fields. The purpose of this paper is to identify the model, or models, with the highest accuracy in predicting a COVID-19 patient's mortality outcome by comparing their performance metrics. Different ML methods employed in model development include logistic regression, decision trees, random forest, eXtreme gradient boosting (XGBoost), multi-layer perceptrons, and the k-nearest neighbors. The metrics used for the comparison of these models were accuracy, precision-recall, F1 score, area under the receiver operating characteristic curve (AUC-ROC), and runtime. The data used comprised the clinical characteristics and histories of 12,425,179 individuals who attended health facilities in Mexico. Following a comprehensive evaluation, the XGBoost model achieved the highest overall score across all metrics. It scored 93.76% in precision, 95.47% in recall, 91.13% in F1-score, 97.86% in AUC-ROC, and had a runtime of 6.67306 s. Therefore, XGBoost was determined to be the preferred method for predicting the mortality outcome of COVID-19 patients.

Keywords: COVID-19; Pandemic; Machine learning; Classification algorithm

***Corresponding author:**
Nikolaos Kourmpanis
(nikos.kourbanis@gmail.com)

Citation: Kourmpanis N, Liaskos J, Zoulias E, Mantas J. Predicting mortality outcomes in individual COVID-19 patients using machine learning algorithms. *Artif Intell Health*. 2024;1(3):31-52. doi: 10.36922/aih.2591

Received: December 30, 2023

Accepted: May 9, 2024

Published Online: July 22, 2024

1. Introduction

Coronavirus disease 2019 (COVID-19)[1] was first identified in the Chinese city of Wuhan in December 2019. On January 30, 2020, the World Health Organization (WHO) classified this outbreak as a Public Health Emergency of International Concern, and on March 11, 2020, it declared COVID-19 a pandemic.[2,3] As of December 13, 2023, there have been 772,386,069 confirmed cases of COVID-19 and 6,987,222 deaths reported to the WHO.[4] COVID-19 is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2),[5] an RNA virus with a spherical shape and a genome composed of a positive-polarity single-stranded RNA.[6] The virus enters human cells through the

angiotensin-converting enzyme 2 (ACE2) receptor. The viral spike protein first attaches to ACE2, and then the membrane enzyme TMPRSS2 cleaves the spike protein, exposing fusion peptides that facilitate fusion with the cell membrane.[7,8] SARS-CoV-2 is the ninth documented coronavirus to infect humans and the seventh identified in the past 20 years.[9,10] Viruses related to SARS-CoV-2 have been documented in bats and pangolins in multiple locations in Southeast Asia, including China, Thailand, Cambodia, and Japan,[11,12] with serological evidence of viral infections in pangolins for more than a decade.[13] Notably, SARS-CoV-2 is primarily transmitted between people through close contact.

The explosion in the number of infections and deaths has led to global efforts to control and curb the disease's spread and associated mortality. One research field that has had a major positive impact on our understanding and fighting the pandemic is machine learning (ML). ML was developed as a tool for data analysis and pattern recognition.[14] ML algorithms process known data and represent it in mathematical ways.[14] During the pandemic, ML studies have assisted in diagnosing and predicting the severity of illness and mortality of COVID-19,[15,16] predicting future mutations of SARS-CoV-2,[17] and promoting the rapid development of therapeutic strategies such as effective vaccines[18] against the virus.

The present study focuses on the prediction of COVID-19 patient mortality using risk factors such as health conditions, habits, and others. Many factors increase the severity of COVID-19 disease, which may even result in the death of the sufferer. A key risk indicator is age, as older people are more likely to get seriously ill from COVID-19. Over 81% of deaths from COVID-19 are among people over the age of 65. The number of deaths among people aged over 65 years is 80 times higher than the number of deaths among people aged 18 – 29 years.[19] There are also medical conditions that increase the severity of the disease, such as heart disease,[19] type I and II diabetes,[20,21] chronic lung diseases,[22] and obesity.[23] In addition, smoking can negatively affect the severity of COVID-19 illness, as it is one of the risk factors for the development and exacerbation of multiple respiratory diseases.[24,25]

The data used in this study were provided by the Epidemiological Surveillance System for Respiratory Diseases under the Directorate-General for Epidemiology of the Ministry of Health of the Government of Mexico.[26] The dataset consisted of over 12 million patients, each with 40 attributes. Many of the attributes, such as geographical data about the patient and health facilities, were dropped as redundant or irrelevant, retaining only those related to pre-existing medical conditions, COVID-19 positivity, and demographics. The remaining attributes were ranked according to their importance scores for each ML method, creating subsets of features with escalating cardinality to be used by different models. Data processing and predictions for each patient's outcome were achieved using models created with six different ML algorithmic methods, namely, logistic regression (LR),[27,28] decision trees (DTs),[29,30] random forest (RF),[31] eXtreme gradient boosting (XGBoost),[32] multi-layer perceptrons (MLPs),[33,34] and the k-nearest neighbors (KNN).[35] The main goal of the present study is to identify the most effective ML method for predicting COVID-19 mortality outcomes with the highest precision,[36] recall,[36] F1 score,[36] and area under the receiver-operator curve (AUC-ROC).[37]

Specifically, this study aims to:

(i). Develop ML models for COVID-19 mortality outcome prediction
(ii). Conduct a comparative analysis of COVID-19 disease mortality outcome prediction using various ML methods (LR, DTs, RF, XGBoost, MLPs, and KNN)
(iii). Evaluate the performance of different ML algorithmic methods used for the prediction of COVID-19 mortality outcome.

2. Related works

This section presents the main characteristics and outcomes of various studies conducted during the COVID-19 pandemic with the aim of predicting the mortality outcome of COVID-19 patients. The data used in these studies varied from purely clinical markers, such as blood test results, to risk factors such as heart disease, obesity, and diabetes included in the patient's history. Sample sizes varied from several hundred to millions. Similarly, the ML methods used in these studies varied, ranging from simple classifiers such as LR, DTs, and KNN to ensemble techniques such as RF, gradient boosting machine (GBM), and XGBoost.

Studies conducted at the beginning of the pandemic that used only one ML method to train the models for predicting COVID-19 patient mortality include Josephus *et al*.[38] who used the LR method in a dataset of 485 patients, and Yan *et al*.[39] who used XGBoost models with a dataset of 1,085 patients. Both studies reported an overall accuracy of 97% for their respective models. However, these studies were limited by their use of only one ML method for the model training and relatively small sample sizes.

Pourhomayoun and Shakibi[40] used a variety of ML methods, including artificial neural networks (ANNs), RF, DTs, support vector machine (SVM), KNN, and LR, to predict mortality in COVID-19 patients. Their

dataset comprised more than 2,670,000 confirmed COVID-19 patients from 146 countries, with an average age of 44.75 years. They applied feature selection to filter irrelevant symptoms and pre-existing conditions, obtaining accuracies of 89.98% for ANNs, 89.83% for KNNs, 89.02% for SVM, 87.93% for RF, 87.91% for LR, and 86.87% for DTs. The advantages of the study include the diversity the patient origins, the large sample size, and the use of different ML methods. The main limitation was the lack of ensemble ML methods.

Naseem *et al.*[41] aimed to develop a novel deep learning neural network (DNN) model for COVID-19 mortality prediction using the Neo-V framework and compared its performance to other traditional ML models, such as LR, RF, KNN, random trees, support vector classifier using radial basis function (SVC-RBF), adaptive boosting (AdaBoost) classifier, quadratic discriminant analysis, and a DNN. The dataset used comprised laboratory and clinical data of 1,214 adult COVID-19 patients admitted to Aga Khan University Hospital from February to September 2020. The DNN Neo-V model outperformed the conventional ML models, achieving an accuracy of 99.53%, a sensitivity of 89.87%, a specificity of 95.63%, and an AUC-ROC of 88.5. The main advantage of the study is the diversity of the ML methods used, including the Neo-V framework, with the limitation being the small number of patients.

Chadaga *et al.*[42] aimed to predict mortality among COVID-19 patients using epidemiological parameters. The ML methods used were RF, XGBoost, LightGBM, categorical boosting, AdaBoost, and gradient boost. The dataset used was provided by the Directorate General of Epidemiology, Secretariat of Health (Mexico)[43] and consisted of 263,007 confirmed COVID-19 patients with 19 selected attributes each. The XGBoost model achieved the best results with an accuracy of 96%, a precision of 95%, a recall of 95%, an F1-score of 95%, and an AUC-ROC of 96%. The advantages of the study include the number and variety of ML methods used and the large patient dataset. However, the main limitation was the lack of non-ensemble ML methods.

Rai *et al.*[44] proposed a voting ensemble model comprising the extra trees classifier, the RF, the gradient boosting classifier, and the XGBoost. The proposed model was compared to baseline models, including KNN, Naïve Bayes classifier, XGBoost, RF, gradient boosting classifier, and extra tree classifier. The dataset used for the research was obtained from a publicly available source consisting of blood biomarkers of 4,711 patients admitted to the hospital from March 1 to April 16, 2020. The highest scores were recorded by the proposed voting ensemble model, with an accuracy of 86.99%, a precision of 0.744, a recall of 0.690, an AUC-ROC of 0.895, and an F1-score of 0.716. The advantage of the study is the use of a diverse ensemble of ML methods, while the main limitation was the small number of patients.

Bárcenas and Fuentes-García[45] conducted a study to determine the risk factors associated with mortality in COVID-19 patients using RF, GBM, and XGBoost. They used a subset of the dataset provided by the Mexican government,[26] recorded from January 17, 2020, to June 28, 2020, which consisted of 583,678 patients, 220,657 of whom were confirmed COVID-19 patients. Patients were classified into three risk categories: low, moderate, and high, depending on comorbidities and major symptoms. The overall accuracy for predicting mortality was 89.97% for XGBoost, 89.86% for RF, and 83.37 for GBM. The advantage of the study is the large patient dataset, while its limitations include the use of only a few ML methods and the lack of non-ensemble methods. In addition, recent studies demonstrate promising results in the use of ML in various domains, such as contact tracing for COVID-19 transmission,[46] prenatal screening,[47] predicting the occurrence of type 2 diabetes,[48] and cardiovascular disease.[49]

3. Data and methods

In this section, we present the dataset used, the preprocessing techniques applied, and the ML algorithmic methods employed to train the models.

3.1. Data preprocessing

The data preprocessing procedure encompasses cleaning, transforming, and encoding the raw data, as well as generating the training and testing datasets for each iteration.

3.1.1. Cleansing

Our dataset consists of 12,425,179 cases suspected of having COVID-19, who attended various health facilities in Mexico from January 17, 2020, until January 3, 2022. The dataset is publicly available as a CSV file disseminated by the Government of Mexico.[26]

First, we translated all 40 attribute names from Spanish to English. Second, we cleansed the dataset by retaining only the positive COVID-19 cases, as indicated by the "Laboratory Result" (1: SARS-CoV-2 positive) and "Final Classification" (1, 2, and 3: Confirmed case) attributes, in accordance with the guidelines of the Epidemiological Association of Mexico and the Mexican Commission of Medical Decisions. Third, we discarded 184,345 cases containing invalid (98: Ignored and 99: Not Specified) or null values in one or more of their attributes. Thus, we

ended up with 3,809,119 COVID-19 patients with valid attribute values. Fourth, we removed the non-correlated attributes – for instance, those related to geography, residency, and indigeneity – resulting in 24 attributes. Finally, we transformed the "Date of Death" attribute into a categorical attribute by renaming it "Survived" and replacing all the "9999-99-99" date values with "1" (Yes) and the rest with "0" (No). We also combined the "Symptom Onset Date" and "Hospital Admission Date" attributes into a numerical attribute labeled "Days from Symptom to Hospitalization." Hence, we settled with 23 attributes, as shown in Table 1. Nineteen of these attributes were included in the ML models, with "Survived" being the target attribute used for predictions, and the remaining four being the "Registration ID" and three indicators. The created "gold standard dataset" (Table 2) consisting of the 23 attribute values of the 3,809,119 patients was saved in CSV file format. The dataset cleansing process flowchart is shown in Figure 1. The dataset value distribution for the categorical attributes is depicted in Figure 2, whereas Figure 3 illustrates the distribution of age groups for both genders. In Figure 3, we observe that most of the patients

Table 1. The 23 attributes of each patient

Attribute name	Values
Registration ID	Patient's unique identification code
Sex	1: Female; 2: Male
Age	Numerical positive
Smoker	1:Yes; 2:No
Pneumonia	1:Yes; 2:No
Diabetes	1:Yes; 2:No
Obesity	1:Yes; 2:No
COPD	1:Yes; 2:No
Asthma	1:Yes; 2:No
Immunosuppressed	1:Yes; 2:No
Hypertension	1:Yes; 2:No
Cardiovascular disease	1:Yes; 2:No
Chronic kidney failure	1:Yes; 2:No
Other chronic disease	1:Yes; 2:No
Pregnancy	1:Yes; 2:No; 97: Not applicable (Male)
Contact with COVID-19 case[a]	1:Yes; 2:No
Laboratory result[a]	1: SARS-CoV-2 positive; 2: SARS-CoV-2 negative; 3, 4: Not clear
Final classification[a,b]	1, 2, 3: Confirmed case; 4: Invalidly identified case; 5, 6, 7: Unconfirmed case
Patient type	1: Not admitted; 2: Admitted
Intubated	1:Yes; 2:No; 97: Not applicable
ICU	1: Admitted to ICU; 2:Not admitted to ICU; 97: Not applicable
Days from symptom to hospitalization[c]	Numerical positive (created attribute)
Survived[b]	1:Yes; 2:No (created attribute)

Notes: [a]Indicators; [b]COVID-19 sample classification; [c]Created attributes.
Abbreviations: COPD: Chronic obstructive pulmonary disease; ICU: Intensive care unit.

Table 2. The golden standard dataset table

Attribute name	Value distribution
Registration ID	3,809,119 unique values
Sex	1,921,058: Female; 1,888,061: Male
Age	Mean: 40.723; Standard deviation: 17.173; Minimum: 0; Maximum: 121
Smoker	250,558: Yes; 3,558,561: No
Pneumonia	395,528: Yes; 3,413,591: No
Diabetes	403,336: Yes; 3,405,783: No
Obesity	448,412: Yes; 3,360,707: No
COPD	31,477: Yes; 3,777,642: No
Asthma	74,682: Yes; 3,734,437: No
Immunosuppressed	23,825: Yes; 3,785,294: No
Hypertension	526,891: Yes; 3,282,228: No
Cardiovascular disease	44,362: Yes; 3,764,757: No
Chronic kidney failure	42,703: Yes; 3,766,416: No
Other chronic disease	60,307: Yes; 3,748,812: No
Pregnancy	29,332: Yes; 1,891,726: No; 1,888,061: Not applicable (Male)
Contact with COVID-19 case	1,519,968:Yes; 2,289,151: No
Laboratory result	3,802,238: SARS-CoV-2 positive; 0: SARS-CoV-2 negative; 6,881: Not clear
Final classification	3,809,119: Confirmed cases; 0: Invalidly identified cases; 0: Unconfirmed cases
Patient type	3,284,671: Not admitted; 524,448: Admitted
Intubated	60,539: Yes; 463,909: No; 3,284,671: Not applicable
ICU	42,095: Admitted to ICU; 482,353: Not admitted to ICU; 3,284,671: Not applicable
Days from symptom to hospitalization	Mean: 3.673; Standard deviation: 3.160; Minimum: -13[a]; Maximum: 43
Survived	3,558,390: Yes; 250729: No

Note: [a]The minimum value is a negative number in the cases where the patient contacted the disease inside the hospital where he was being treated.
Abbreviations: COPD: Chronic obstructive pulmonary disease; ICU: Intensive care unit.

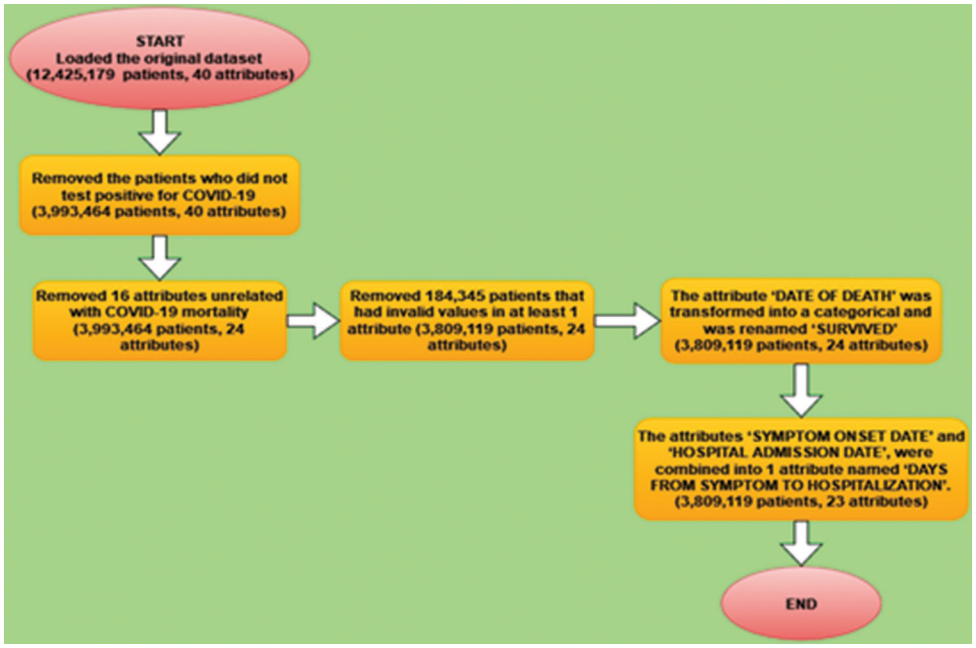


Figure 1. Data cleansing process flowchart. Image created using Draw.io (https://app.diagrams.net/)

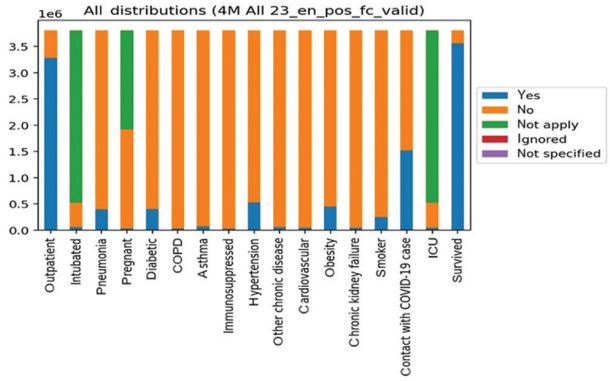


Figure 2. Distribution plot of categorical attribute values in the valid COVID-19 dataset. Image created using Python's Matplotlib library. Abbreviations: COPD: Chronic obstructive pulmonary disease; ICU: Intensive care unit.

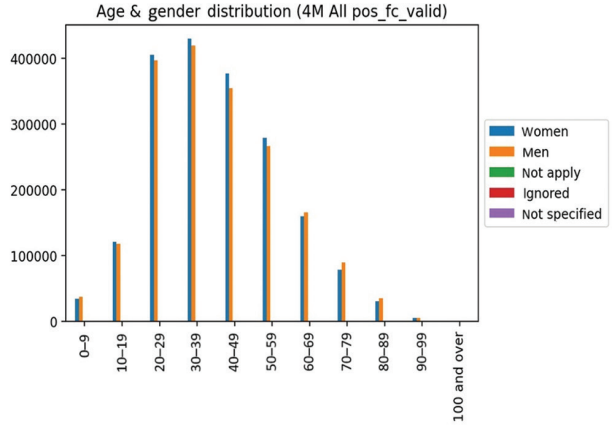


Figure 3. Age group and gender distribution chart in the valid COVID-19 dataset. Image created with Python's Matplotlib library

are between 20 and 59 years old and that the two genders are equally divided across all age groups.

3.1.2. Transformation-encoding

We further encoded and transformed the data values of the newly constructed dataset using Python's statistical analysis libraries and methods. The first step was to encode 14 out of the 17 categorical attributes, excluding "Pregnancy," "Intubated," and "ICU," using sklearn's "LabelEncoder" method. This method assigns a unique value between 0 and n-1 to each distinct value of an attribute, where n is the number of distinct values for that attribute. Next, we used the "get_dummies" method from the pandas library to convert the categorical attributes "Pregnancy," "Intubated," and "ICU" into dummy-pointer variables. These three attributes can take three distinct values: "1" for "Yes," "2" for "No," and "97" for "Not Applied." The "get_dummies" method breaks each attribute into m-1 attributes, where m is the number of distinct values for that attribute. For instance, the "Pregnancy" attribute was split into two new attributes: "Pregnancy_2" and "Pregnancy_97." Here, the values "1,0" describe a non-pregnant female, "0,1" describe a male, and "0,0" describe a pregnant female. The same transformation was applied to the "Intubated" and "ICU" attributes, resulting in "Intubated_2," "Intubated_97,"

"ICU_2," and "ICU_97." After this processing, we created three more categorical attributes, bringing the total to 20. After adding the two numerical attributes "Age" and "Days from Symptom to Hospitalization," we ultimately settled on a final total of 22 attributes used by the ML models. Finally, using the sklearn's statistical methods, namely "StandardScaler" (std) and "MinMaxScaler" (mm), we created six distinct datasets with different normalization schemes for the numerical attributes "Age" and "Days from Symptom to Hospitalization." Specifically, one dataset was created using the "StandardScaler" method, four using the "MinMaxScaler" method with different ranges (0 – 1, 0 – 10, 0 – 100, and 0 – 1000), and one without any normalizing method (none). Each of the six datasets was saved in CSV file format. The transformation-encoding process flowchart is shown in Figure 4.

3.1.3. Train-test set generation

To create the train and test sets for each ML model iteration, we formed a dataset by randomly selecting 20% of the samples from the current dataset file using the "sample" function from the pandas library. Next, to mitigate the imbalance in the "Survived" attribute, which had an approximate dead-to-survivor ratio of 1:14, we applied the synthetic minority oversampling technique (SMOTE)[50] from Python's imblearn library. This adjustment created a set with a dead-to-survivor ratio of 1:10. Finally, we applied the "RandomUnderSampler" method from the imblearn library to create the final dataset with a dead-to-survivor ratio of 1:2. This dataset was then randomly divided into two subsets: the train set, consisting of 70% of the data, and the test set, consisting 30% of the data. The flowchart illustrating the process of generating the train and test sets is depicted in Figure 5.

3.2. Models and algorithmic methods

In this study, we used six ML algorithmic methods, i.e., LR,[27,28] DTs,[29,30] RF,[31] XGBoost,[32] MLPs,[33,34] and KNN.[35] The models were implemented in Python (version 3.7) using the integrated development environment (IDE) software Spyder (version 5.1.5) and the pandas library (version 1.3.5).

3.2.1. LR

LR is a supervised learning method developed by David Cox in 1958[27] that aims to solve classification problems. LR is a generalized form of simple linear regression, used for solving classification problems where both numerical variables and categorical variables can be used as dependent variables. LR models data using the sigmoid function to make predictions about different possible outcomes.[28] Specifically, it predicts the value of dependent variables based on the weights of each independent variable. The weight of each variable is related to the degree of correlation it has with the dependent variable. In this study, we used the "LogisticRegression" method from Python's sklearn library.

3.2.2. DTs

DTs are a non-parametric supervised learning method used for classification or regression problems. The main goal of DTs is to build a model that predicts the value of a target variable by learning simple decision rules inferred from data features. The algorithm takes the given dataset and divides it into categories consisting of entities with the same value for a specific variable (attribute). This process is repeated recursively until the DT is constructed through the rules of the individual categorizations of the specific model.[30] A tree can be thought of as a piecewise consistent approximation. In this study, we constructed our DT models using the "DecisionTreeClassifier" method from Python's sklearn library.

3.2.3. RF

RF is an ensemble-supervised ML method that can be applied to both classification and regression problems. RF improves model performance by combining multiple classifiers to solve complex problems.[31] Specifically, RF is a classifier that consists of a number of DTs, each trained on a different subset of the training set. The final decision (prediction) is made by the majority vote for categorical variables or by averaging the values for numerical variables, enhancing the model's accuracy. The higher the number of DTs that comprise the forest, the higher the model's accuracy and the lower the risk of overfitting. In this study, we used the "RandomForestClassifier" method from the sklearn library.

3.2.4. XGBoost

XGBoost is a well-known variant of the gradient boosting algorithm, developed to increase prediction accuracy. XGBoost is an ensemble learning method based on DTs, utilizing a gradient-boosting framework. This framework corrects mistakes from previous DT models by modifying the weights of the variables, thereby improving subsequent models. The method was originally developed by Tianqi Chen and described by him and Carlos Guestrin in 2016.[32] XGBoost has gained widespread popularity due to its performance in ML competitions.[32] In this study, we used the "XGBClassifier" method from Python's XGBoost library.

3.2.5. MLPs

MLPs are another term for ANNs since the artificial neuron is also called a "Perceptron."[51] MLPs are a

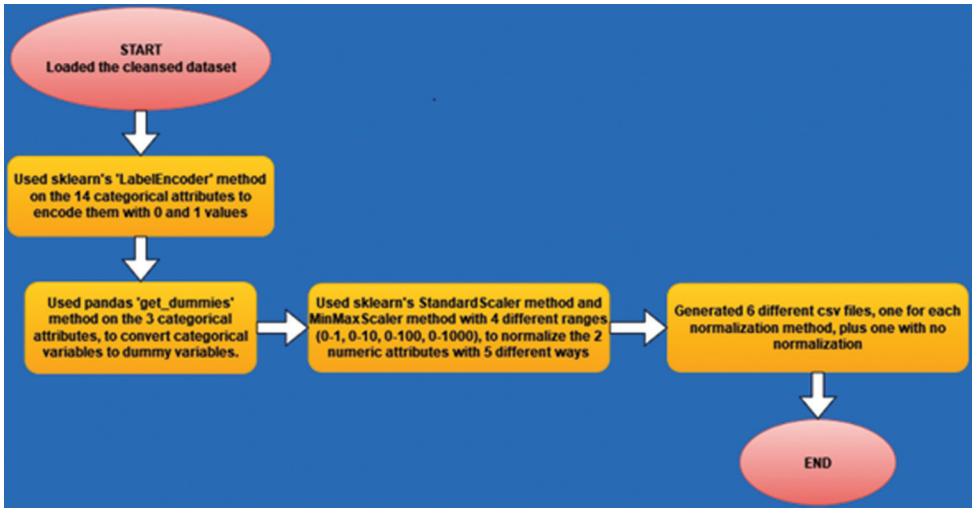


Figure 4. Data transformation-encoding process flowchart. Image created using Draw.io (https://app.diagrams.net/)

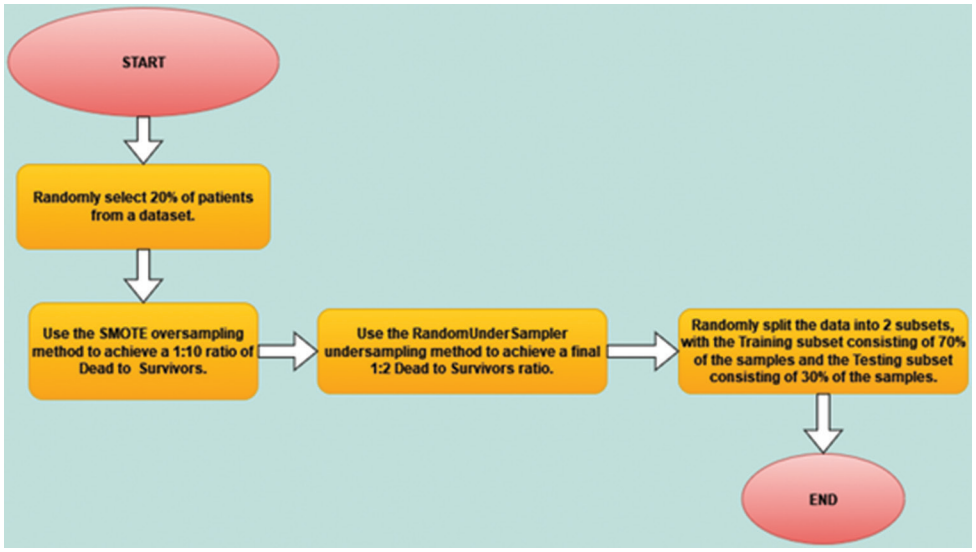


Figure 5. Train-test set generation process flowchart. Image created using Draw.io (https://app.diagrams.net/).
Abbreviation: SMOTE: Synthetic minority oversampling technique.

supervised learning method that mimics the function of neurons through algorithmic techniques to solve complex computational problems.[33,34] Given a set of attributes as independent variables and one independent variable as the target, MLPs can "learn" a non-linear approximation of the function for either classification or regression. MLPs consist of several layers of neurons: the input layer, the output layer, and the in-between or hidden layers. Each neuron is connected to every neuron of the previous and next layers. The input layer's neurons receive the data used to make predictions and pass it to the hidden layers. Finally, the output layer's neurons receive the values from the last hidden layer and make predictions for the corresponding classification or regression problem. As more layers are added to ANNs, the gradients of the loss function approach zero, making the network hard to train due to the vanishing gradient problem. This problem can be overcome through a multi-pronged approach, varying from the utilization of rectified linear unit activations to new algorithms exploring fresh techniques or enhancing existing ones.[52] MLPs are capable of learning non-linear models in real-time, but, due to the hidden layers, they exhibit a non-convex loss function owing to multiple local minima. Therefore, different random attribute weight initializations can lead to different validation accuracies, as MLPs are sensitive to the scaling of attribute weights.[53] In the present study, we used the "MLPClassifier" method from the sklearn library.

3.2.6. KNNs

The KNNs are a non-parametric supervised learning method used in classification problems, where "non-parametric" means that the input and output data will be similar in type. The method was discovered by Fix and Hodges[35] in 1951 and was subsequently developed by Cover.[54] KNN classifies new samples based on their value distance from samples with a known class label, relying on the logic that similar samples belong to the same class.[55] The class to which each new sample will belong depends on its distance from the k previous samples in the training dataset. KNN can be used for classification problems with discrete variable objectives or regression problems with continuous variable objectives. In this study, we used the "KNeighborsClassifier" method from the sklearn library.

3.3. The importance of attributes

For each ML method, except for MLPs and KNN, we used three different sets of attributes, depending on the importance score that each attribute aggregated according to the "feature_importances_" method. This sklearn library method is a vector of shape available in certain Python predictors and provides a relative measure of the importance of each feature in the predictions of the model.[56] For the "MLPClassifier" and "KNeighborsClassifier," the score for each attribute was calculated as the normalized sum of the scores from the four previous methods: "LogisticRegression," "DecisionTreeClassifier," "RandomForestClassifier," and "XGBClassifier."

These three sets had a different number of attributes: One contained all 22 attributes, another included the 15 most important attributes and the last contained only the 10 most important attributes. The following diagrams in Figures 6-16 illustrate the attribute rankings and the SHAP (SHapley Additive exPlanation) summary plots for all six ML methods.

3.4. Hyperparameter values optimization

We used three different sets of hyperparameters for each ML method. The first set contained the default values (default), the second set contained the first set of optimized values for the ML method's hyperparameters (opt_01), and the third set contained the second set of optimized hyperparameters values (opt_02). These two optimized hyperparameter sets were created using the "GridSearchCV" method from the sklearn library. To form the two sets of optimized hyperparameters, including the optimal values for most hyperparameters, we applied sklearn's "GridSearchCV()" grid search method. This method is used to search for the optimal value of each hyperparameter through a given grid containing all possible values of the different hyperparameters. In this study, we used "GridSearchCV()" as a 10-fold cross-validation method. It accepts the respective ML method and the sets of hyperparameter values as input and outputs the optimal value for each hyperparameter. This process resulted in nine different combinations for every ML method across the six datasets, creating a total of 54 different models for each ML method, and 324 models in total for all six methods. We ran each model 10 times (iterations) and calculated the mean to avoid extreme values in the metrics, resulting in 540 iterations for each ML method and a total of 3,240 iterations for all ML methods. The flowchart for creating, training, and evaluating each model is shown in Figure 17.

4. Results

This section presents the metrics used in the evaluation and the evaluation results of the created models. The evaluation results are presented for both all models created as well as the models with the highest overall score for each ML method. In addition, an overall ranking of all models according to their highest score is presented.

4.1. Evaluation metrics

To assess the performance of 324 ML models, we used the metrics of precision,[36] recall,[36] F1 score,[36] and the AUC-ROC,[37] computed through the confusion matrix,[36] and the runtime metric. The confusion matrix is a summary of the prediction results of a model, depicting the number of correct and incorrect predictions made by the evaluation model. The predictions are categorized into four groups: True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).[36] The TP is the correctly predicted positive value, FP is the wrongly predicted positive value, TN is the correctly predicted negative value, and FN is the wrongly predicted negative value for the samples of the training set. Based on these four parameters, we can calculate precision, recall, F1 Score (F1 Score), and the AUC-ROC.

Precision is calculated as the ratio of TP to the total predicted positive observations, giving us the model's percentage of correctly predicted positive values. It is given by Equation I.

$$\text{Precision}=\frac{TP}{(TP+FP)} \quad \text{(I)}$$

Recall is the ratio of TP to the total number of positive values. It is given by Equation II below.

$$\text{Recall}=\frac{TP}{(TP+FN)} \quad \text{(II)}$$

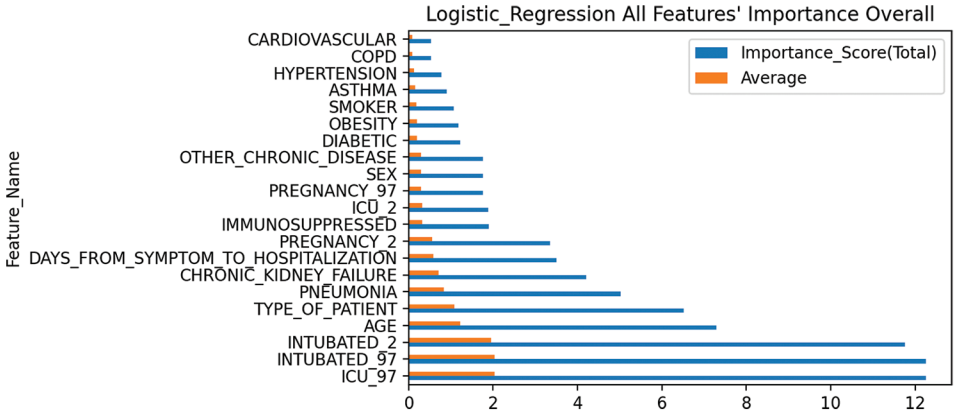


Figure 6. Attribute importance ranking of the "LogisticRegression" method. Image created using Python's Matplotlib library

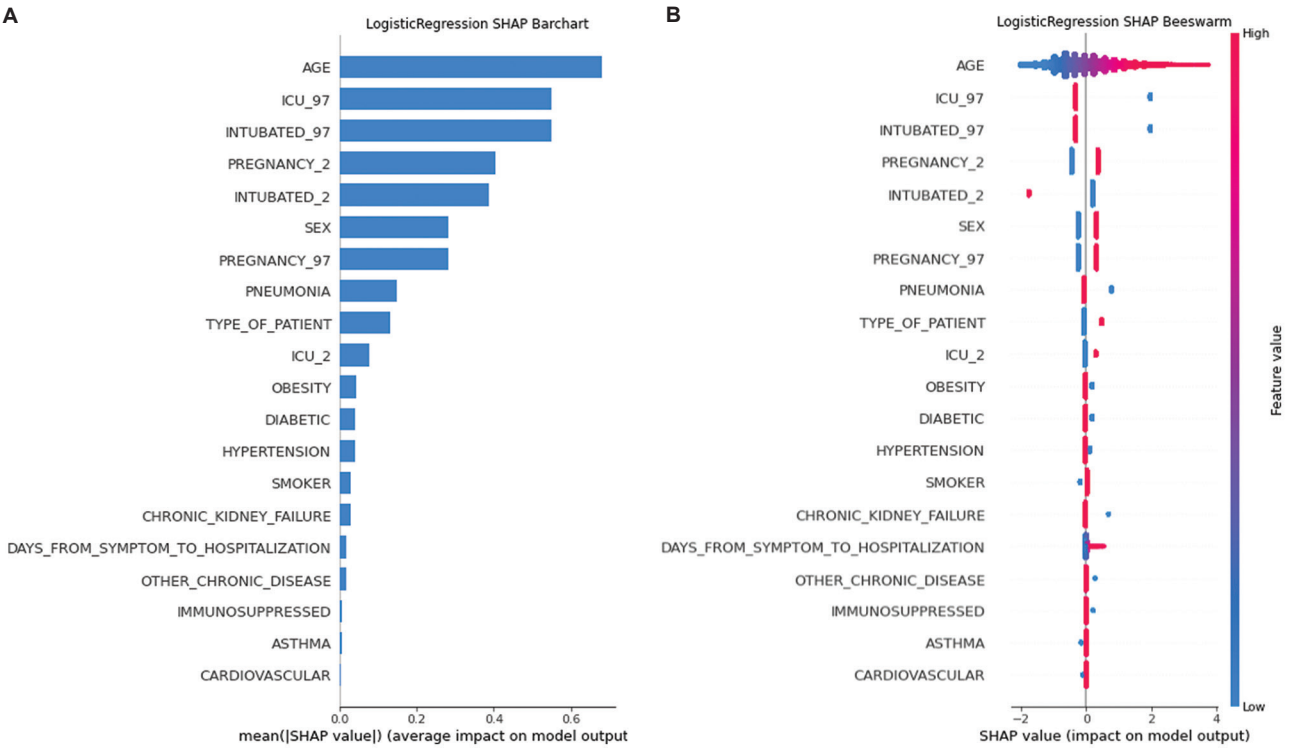


Figure 7. SHAP summary plots of the "LogisticRegression" method. (A) Barchart. (B) Beeswarm. Image created using Python's Matplotlib library

The F1 score is the weighted average of precision and recall, taking into account both FP and FN. It is usually more useful than precision, especially if there is an uneven target class distribution. The F1 score computation is given by Equation III.

$$\text{F1 score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{(\text{Recall} + \text{Precision})} \quad \text{(III)}$$

The AUC-ROC is calculated as the entire two-dimensional area under the receiver operating characteristic (ROC) curve (Figure 18), from 0.0 to 1.1.[37] The ROC curve depicts the performance of the ML model being evaluated across all classification thresholds. Specifically, the ROC curve is a representation of the true positive rate (TPR) and false positive rate (FPR).[37] As the classification threshold is lowered, the model classifies more items as positive, resulting in an increase for both FPs and TPs. The value of AUC-ROC ranges from 0 to 1; for example, for a model with 100% inaccurate predictions, the AUC-ROC will be 0.00, whereas for a model with 100% accurate predictions, the AUC-ROC will be 1.00.[37] TPR and FPR are calculated using Equations IV and V, respectively.

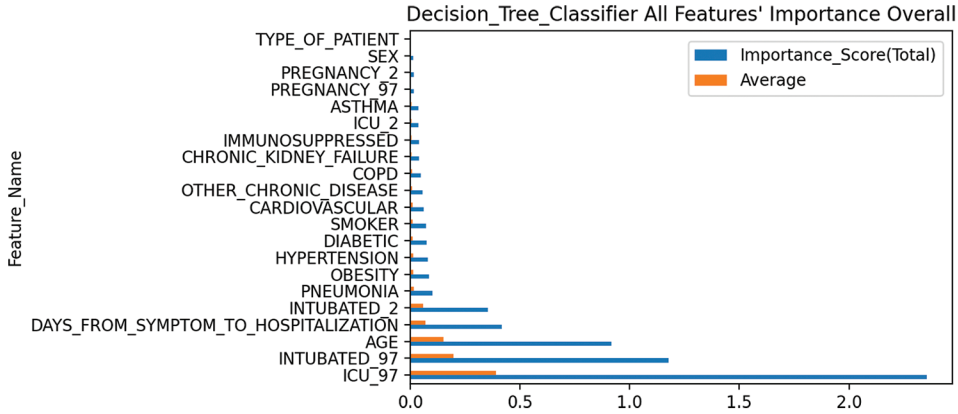


Figure 8. Attribute importance ranking of the "DecisionTreeClassifier" method. Image created using Python's Matplotlib library

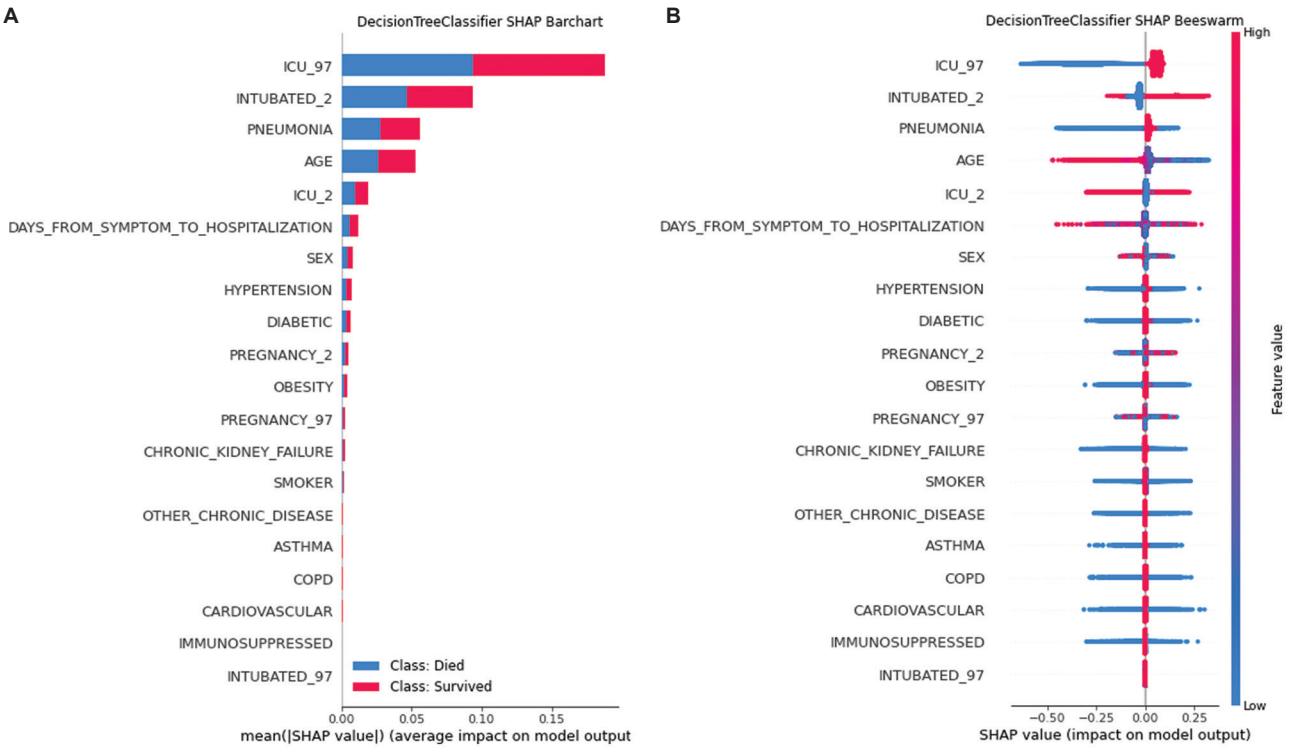


Figure 9. SHAP summary plots of the "DecisionTreeClassifier" method. (A) Barchart. (B) Beeswarm. Image created using Python's Matplotlib library

$$TPR=\frac{TP}{(TP+FN)} \quad (IV)$$

$$FPR=\frac{FP}{(FP+TN)} \quad (V)$$

Runtime is an additional metric that we used; it is measured in seconds and defined as the duration of a model's iteration.

4.2. Model evaluation

After running and evaluating all 324 different models, we ranked them according to their scores. The ML models achieved precision ranging from 90.09% to 93.76%, recall from 83.42% to 96.99%, F1-score from 84.96% to 91.13%, AUC-ROC from 0.9003 to 0.9788, and runtime from 1.092 to 910.173 s. The results of this ranking are depicted in Figure 19. The model with the highest score is positioned on the leftmost position, with the metric values decreasing as we move toward the right, so the last model scores the

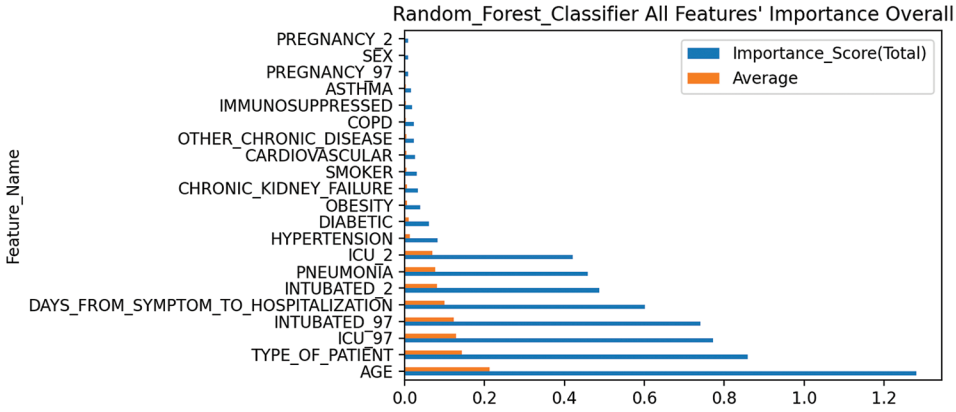


Figure 10. Attribute importance ranking of the "RandomForestClassifier" method. Image created using Python's Matplotlib library

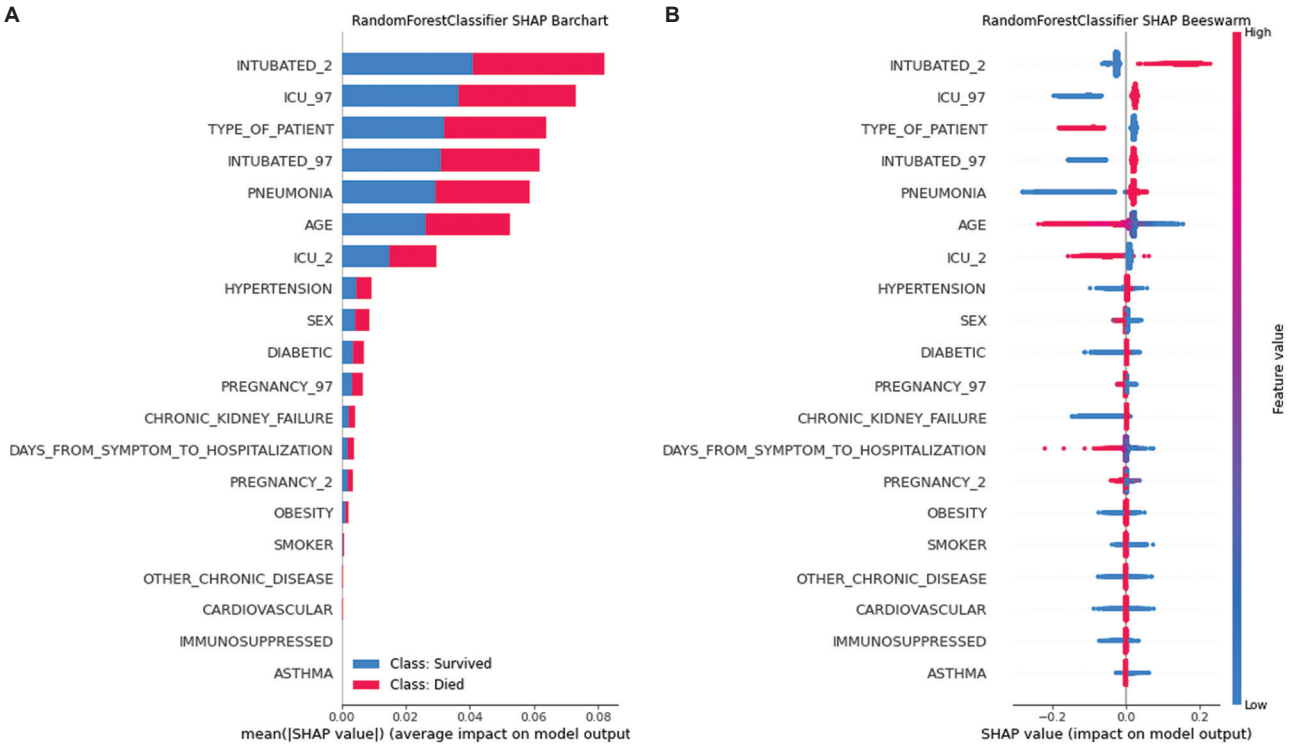


Figure 11. SHAP summary plots of the "RandomForestClassifier" method. (A) Barchart. (B) Beeswarm. Image created using Python's Matplotlib library

lowest. In the following sections, we analyze and rank the performance of the different ML methods' models both for each metric and overall for all metrics, presenting each method's highest-performing model. The ranges of the values of each metric for all models and for the three highest-ranking models for each ML method are illustrated in Tables 3 and 4, respectively. All appendix files are publicly available on GitHub (https://github.com/NikosKourb/Patients_Mortality_COVID-19_ML). All the model processes described in this paper were run in a Spyder 5.3.3 version IDE using Python 3.7.1 version as the programming language in a Microsoft Windows 10 Enterprise (x64) Build 19045.3570 (22H2) environment. The hardware used was a DELL Inspiron 3576 laptop, with an Intel(R) Core(TM) i7-8550U CPU (4 cores/8 threads/1.80GHz base/4.00GHz max), 8GB of DDR4 (2400/PC4-19200/1200.0 MHz) SDRAM, an Intel UHD Graphics 620 (Kaby Lake R U GT2) video adapter, and a DELL 0J11DH motherboard.

4.2.1. Precision

The XGBoost models showed the highest precision values, ranging from 93.21% (113th position) to 93.76%

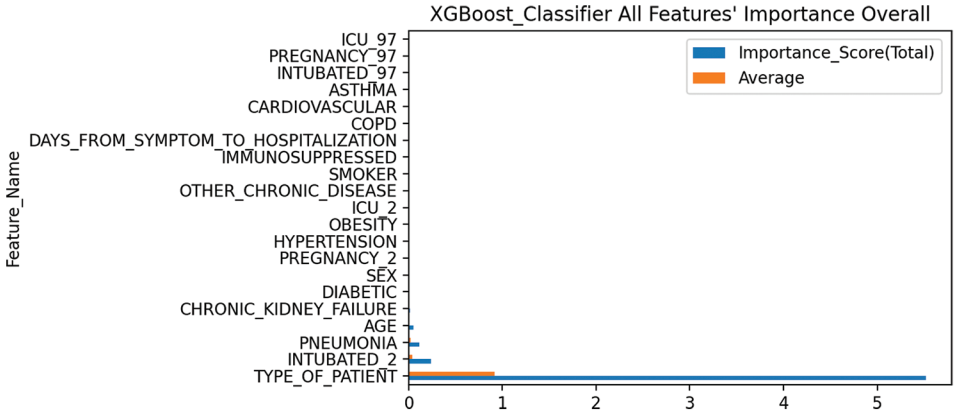


Figure 12. Attribute importance ranking of the "XGBClassifier" method. Image created using Python's Matplotlib library

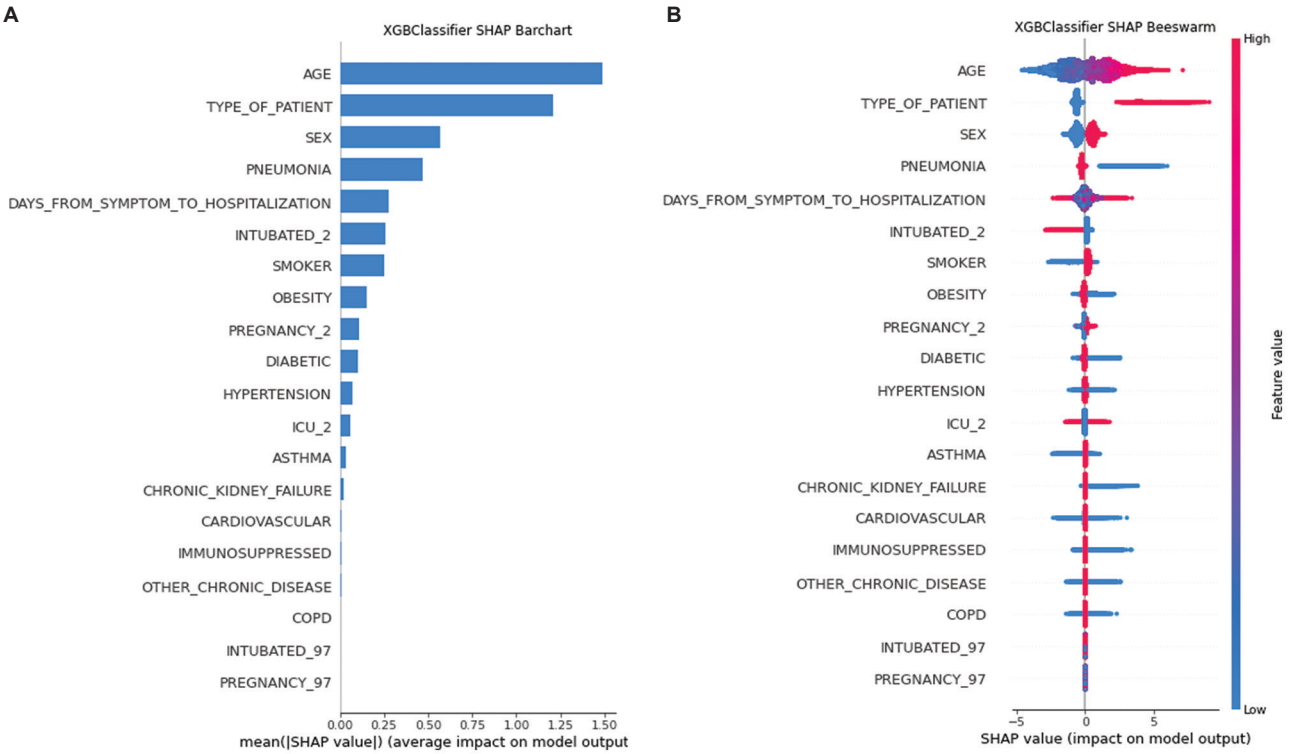


Figure 13. SHAP summary plots of the "XGBClassifier" method. (A) Barchart. (B) Beeswarm. Image created using Python's Matplotlib library

(1st position). Half of the XGBoost models (28/54) ranked above the 55th place. The highest-ranked XGBoost models processed datasets using the "Min–Max" method, with ranges of 0 – 100 and 0 – 1000 (mm_0 – 100 and mm_0 – 1000, respectively) and used all 22 attributes with the first set of optimized (opt-01) hyperparameter values.

Following XGBoost, the RF models exhibited precision values ranging from 92.25% (273rd position) to 93.66% (11th position). The distribution of RF model rankings demonstrated significant dispersion, with 50% of the models ranking above the 146th position. The highest-ranked RF models processed datasets with the "Min–Max"" method, using ranges of 0 – 100 and 0 – 1000 (mm_0 – 100 and mm_0 – 1000, respectively), and used either 22 or 15 attributes with the second set of optimized (opt-02) hyperparameter values.

The MLP models ranked third, with precision scores ranging from 92.59% (172nd position) to 93.46% (32nd position), with half of them ranking above the 88th position. The highest-ranked MLP models processed datasets with the "Min–Max" method, using ranges of

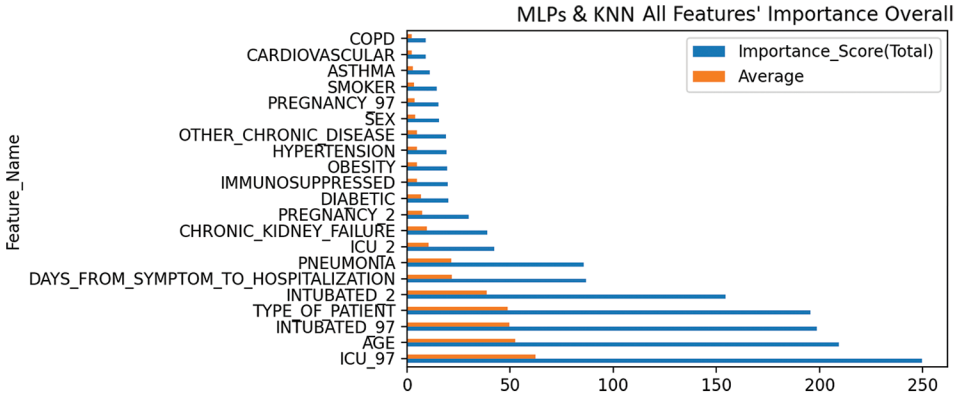


Figure 14. Attribute importance ranking of "MLPClassifier" and "KNeighborsClassifier" methods. Image created using Python's Matplotlib library

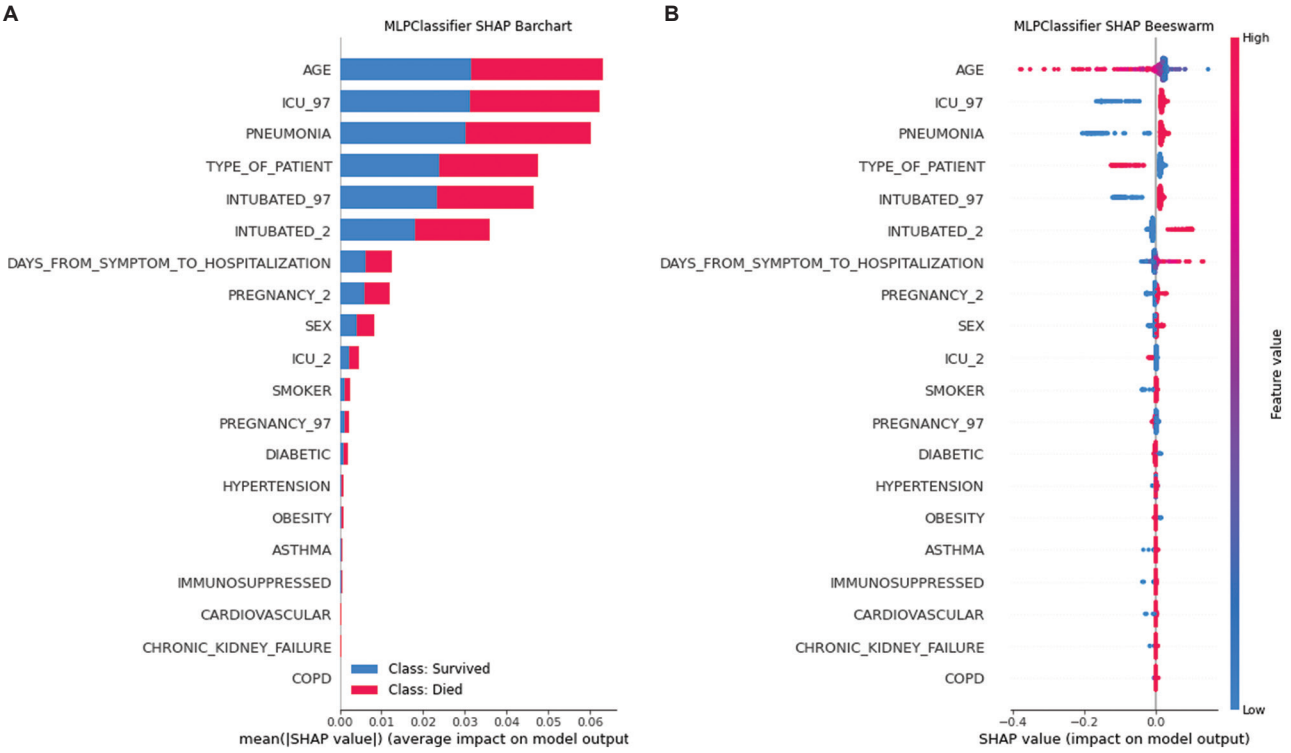


Figure 15. SHAP summary plots of the "MLPClassifier" method. (A) Barchart. (B) Beeswarm. Image created using Python's Matplotlib library

0 – 100 and 0 – 1000 (mm_0 – 100 and mm_0 – 1000, respectively), and used 22 attributes with either the first or the second sets of optimized (opt-01 and opt-02, respectively) hyperparameter values.

In the fourth place were the DT models, with precision scores ranging from 90.09% (324th position) to 93.04% (127th position). The highest-scoring DT models handled datasets that were either not processed with any normalization method (none) or processed with the "Min–Max" method, with a range of 0 – 1000 (mm_0 – 1000), and used either 22 or 15 attributes with the first set of optimized (opt-01) hyperparameter values.

K-nearest neighbor models ranked fifth, with precision scores ranging from 91.54% (304th position) to 92.85% (142nd position). The highest-scoring KNN models used datasets processed with the "StandardScaler" method (std), used either 22 or 15 attributes, and employed the first set of optimized (opt-01) hyperparameter values.

Finally, the LR models showed precision values ranging from 92.32% (267th position) to 92.63% (169th position).

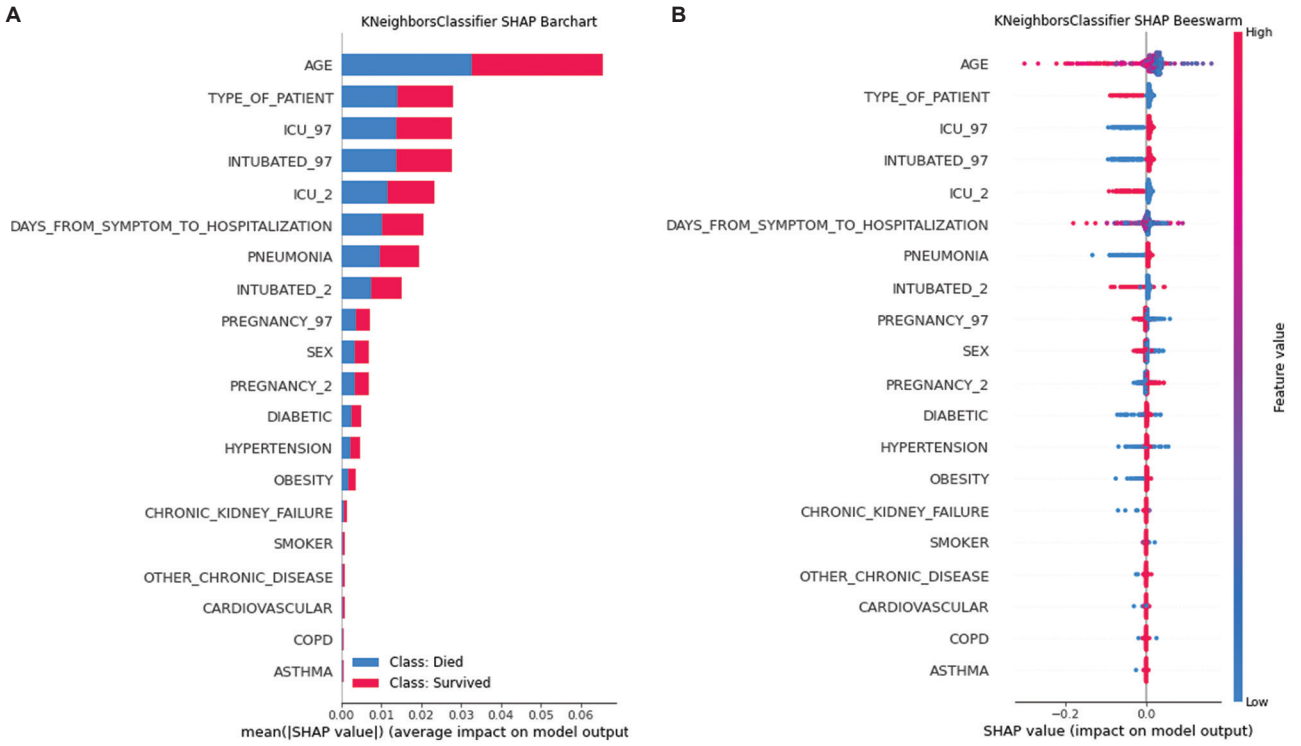


Figure 16. SHAP summary plots of the "KNeighborsClassifier" method. (A) Barchart. (B) Beeswarm. Image created using Python's Matplotlib library

The highest-ranked LR models processed datasets processed using the "Min–Max" method, with ranges of 0 – 100 and 0 – 1000 (mm_0 – 100 and mm_0 – 1000, respectively), and used either 22 or 15 attributes with the default set of (default) hyperparameter.

4.2.2. Recall

The RF models showed the highest values for recall, ranging from 90.83% (271th position) to 96.99% (1st position). The distribution of the model's rankings demonstrates significant dispersion, with 33.3% (18/54) ranking above the 19th position and 44.4% (24/54) occupying positions between 119th and 176th. The highest-ranked RF models handled datasets processed with the "Min–Max" method, using ranges of 0 – 100 and 0 – 1000 (mm_0 – 100 and mm_0 – 1000, respectively), and used either 22 or 15 attributes with the second set of optimized (opt-02) hyperparameter values.

Following the RF models, the XGBoost models ranked second, with recall values ranging from 93.96% (133rd position) to 95.61% (20th position), and 63% (34/54) ranking above the 70th position. The highest-ranked XGBoost models processed datasets using the "Min–Max" method, with ranges of 0 – 100 and 0 – 1000 (mm_0 – 100 and mm_0 – 1000, respectively), and used either 22 or 10 attributes with either the default or the first set of optimized (opt-01) hyperparameter values.

In third place were the MLPs models, scoring from 93.43% (142nd position) to 95.62% (19th position), with 88.9% (48/54) occupying positions between 51st and 113th. The highest-ranked MLPs models processed datasets using the "Min–Max" method, with ranges of 0 – 100 and 0 – 1000 (mm_0 – 100 and mm_0 – 1000, respectively) and used either 22 or 10 attributes with either the first or the second sets of optimized (opt-01 and opt-02, respectively) hyperparameter values.

In the fourth place are the DTs models, scoring from 83.42% (324th position) up to 93.69% (135th position). The highest scoring DT models handled datasets that were either not processed with any normalization method (none) or were processed with the "Min–Max" method, with 0 – 1000 (mm_0 – 1000) range, used either 22 or 15 attributes and the first set of optimal (opt-01) hyperparameter values.

In fifth place were the KNN models, scoring from 88.23% (303rd position) to 92.95% (157th position). The highest-ranked KNN models used datasets processed with the "StandardScaler" method (std) and used either 22 or 15 attributes with the first set of optimized (opt-01) hyperparameter values.

Lastly, the LR models showed recall values ranging from 90.79% (278th position) to 91.89% (182nd position). The highest-ranked LR models processed datasets using

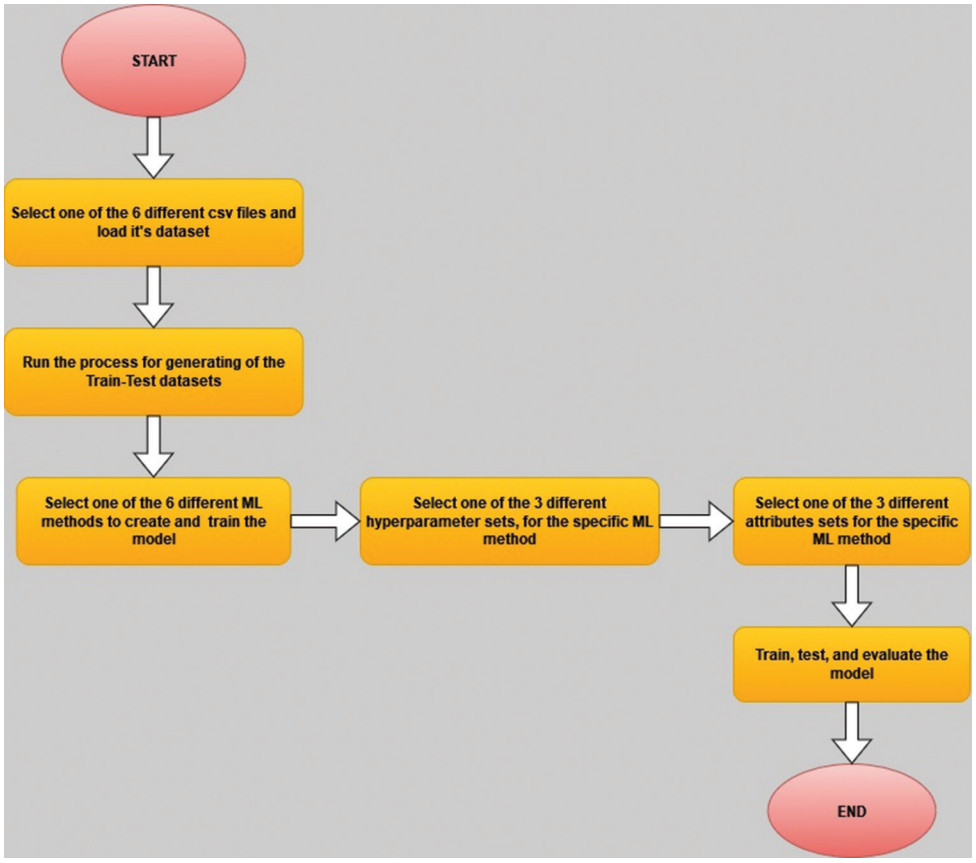


Figure 17. Creation, training, and evaluation process flowchart for each of the 324 models. Image created using Draw.io (https://app.diagrams.net/). Abbreviation: ML: Machine learning.

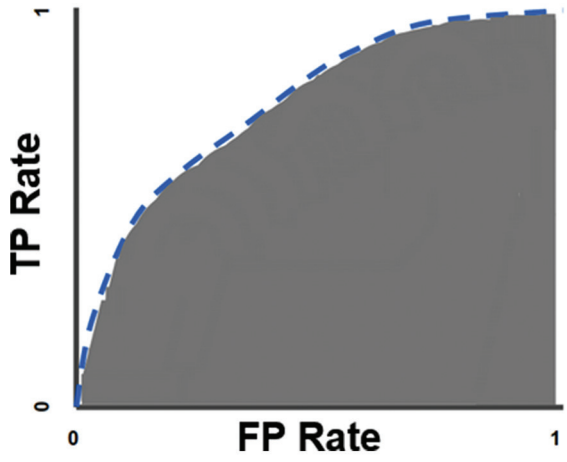


Figure 18. The area under the receiver-operating characteristic curve.[37] Abbreviations: FP: False positive; TP: True positive.

the "Min–Max" method, with ranges of 0 – 1 and 0 – 1000 (mm_0 – 1 and mm_0 – 1000, respectively) and used either 22 or 10 attributes with either the first or the second sets of optimized (opt-01 and opt-02, respectively) hyperparameter values.

4.2.3. F1 score

The XGBoost models demonstrated the highest F1 scores, ranging from 90.33% (121st position) to 91.13% (1st position), with 63% (34/54) ranking above the 54th position. The highest-ranked XGBoost models processed datasets using the "Min-Max" method, with ranges of 0 – 10, 0 – 100, and 0 – 1000 (mm_0 – 10, mm_0 – 100, and mm_0 – 1000, respectively), used 22 attributes, and employed the first set of optimized (opt-01) hyperparameter values.

The RF models secured second place, with values ranging from 88.73% (274th position) to 91.13% (2nd position). The highest-ranked RF models handled datasets processed with the "Min–Max" method, with ranges of 0 – 100 and 0 – 1000 (mm_0 – 100 and mm_0 – 1000, respectively), used either 22 or 15 attributes, and utilized either the default or the second set of optimized (opt-02) hyperparameter values.

In the third place were the MLPs models, which scored from 89.39% (168th position) to 90.76% (34th position). The highest-ranked MLP models used datasets processed

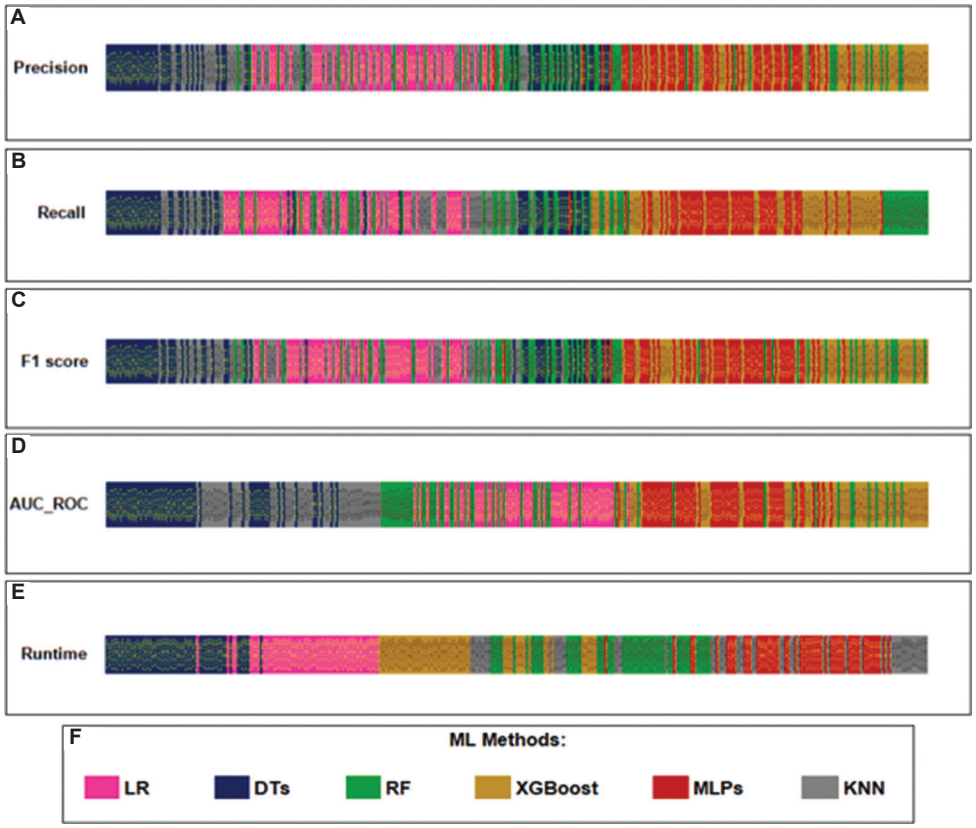


Figure 19. Ranking all 324 different models for the different metrics. (A) Ranking for precision. (B) Ranking for recall. (C) Ranking for F1 score. (D) Ranking for AUC-ROC. (E) Ranking for runtime. (F) Color-coding matching for the different ML methods. Image created using Microsoft Excel. Abbreviations: AUC-ROC: Receiver operating characteristic curve; DTs: Decision trees; KNN: K-nearest neighbors; LR: Linear regression; ML: Machine learning; MLPs: Multi-layer perceptrons; RF: Random forest; XGBoost: eXtreme gradient boosting.

with the "Min–Max" method, with ranges of 0 – 100 and 0 – 1000 (mm_0 – 100 and mm_0 – 1000, respectively), used either 22 or 15 attributes, and employed either the first or the second sets of optimized (opt-01 and opt-02, respectively) hyperparameters.

The DT models ranked fourth, scoring from 84.96% (127th position) to 90.03% (324th position). The highest-scoring DT models handled datasets that were either not processed with any normalization method (none) or were processed with the "Min–Max" method, with the 0 – 1000 (mm_0 – 1000) range, used either 22 or 15 attributes, and employed the first set of optimized (opt-01) hyperparameter values.

Following the DTs models, in fifth place, were the KNN models that scored from 87.52% (304th position) to 89.74% (145th position). The highest-scoring KNN models used datasets processed with the "StandardScaler" method (std), used either 22 or 15 attributes, and employed the first set of optimized (opt-01) hyperparameter values.

In the last place were the KNN models, with values ranging from 88.88% (266th position) to 89.26% (172nd position). The highest-ranking KNN models handled datasets processed with the "Min–Max" method, with ranges of 0 – 1 and 0 – 1000 (mm_0 – 1 and mm_0 – 1000, respectively), used all 22 attributes, and employed either the default or the second sets of optimized (opt-02) hyperparameter values.

4.2.4. Area under the receiver operating characteristic curve

The XGBoost models showed the highest values for AUC-ROC, with values ranging from 97.07% (122nd position) to 97.88% (1st position). A significant portion, 74.1% (40/54), of the XGBoost models ranked above the 58th position. The top-performing XGBoost model's processed datasets processed using the "Min–Max" method, with ranges of 0 – 100 and 0 – 1000 (mm_0 – 100 and mm_0 – 1000, respectively), used all 22 attributes, and employed either the default or the first set of optimized (opt-01) hyperparameter values.

The RF models secured second place, with AUC-ROC values ranging from 96.16% (216th position) to 97.71% (11th position). The distribution of the RF models showed a

Table 3. Performance results of all models for each ML method

ML methods	Metrics				
	Precision	Recall	F1-score	AUC-ROC	Runtime (sec)
LR	92.32 – [a]92.63%	90.79 – [a]91.98%	88.88 – [a]89.26%	0.9646 – 0.9708	1.343 – 3.317
DTs	[a]90.09 – 93.04%	[a]83.42 – 93.69%	[a]84.96 – 90.03%	[a]0.9003 – [a]0.9567	[a]1.092 – [a]1.722
RF	92.25 – 93.66%	90.83 – [b]96.99%	88.73 – 91.12%	0.9616 – 0.9771	11.951 – 30.221
XGBoost	[b]93.21 – [b]93.76%	[b]93.96 – 95.61%	[b]90.33 – [b]91.13%	[b]0.9707 – [b]0.9788	4.969 – 17.711
MLPs	92.59 – 93.46%	93.43 – 95.62%	89.39 – 90.76%	0.9706 – 0.9735	[b]18.144 – 362.466
KNN	91.54 – 92.85%	88.23 – 92.95%	87.52 – 89.74%	0.9415 – 0.9593	8.060 – [b]910.173

Notes: [a]Lowest metric values; [b]Highest metric values.
Abbreviations: AUC-ROC: Receiver operating characteristic curve; DTs: Decision trees; KNN: K-nearest neighbors; LR: Linear regression; ML: Machine learning; MLPs: Multi-layer perceptrons; RF: Random forest; XGBoost: eXtreme gradient boosting.

Table 4. Performance results of the three models, with the highest overall scores, for each ML method

ML methods	Metrics				
	Precision	Recall	F1-score	AUC-ROC	Runtime (sec)
LR	[a]92.56 – [a]92.63%	[a]90.99 – [a]91.29%	[a]89.14 – [a]89.26%	0.9704 – 0.9708	2.994 – 3.235
DTs	92.98 – 93.04%	93.37 – 93.69%	89.94 – 90.03%	[a]0.9549 – [a]0.9567	[a]1.147 – [a]1.451
RF	93.55 – 93.66%	[b]96.63 – [b]96.99%	90.96 – [b]91.13%	0.9744 – 0.9771	26.678 – 29.847
XGBoost	[b]93.73 – [b]93.76%	95.40 – 95.61%	[b]91.09 – 91.12%	[b]0.9778 – [b]0.9788	6.128 – 6.741
MLPs	93.40 – 93.46%	95.28 – 95.62%	90.67 – 90.76%	0.9726 – 0.9729	117.228 – 181.845
KNN	92.71 – 92.85%	92.56 – 92.95%	89.50 – 89.74%	0.9583 – 0.9593	[b]48.215 – [b]882.944

Notes: [a]Lowest metric values; [b]Highest metric values.
Abbreviations: AUC-ROC: Receiver operating characteristic curve; DTs: Decision trees; KNN: K-nearest neighbors; LR: Linear regression; ML: Machine learning; MLPs: Multi-layer perceptrons; RF: Random forest; XGBoost: eXtreme gradient boosting.

significant dispersion, with the lowest half ranking between the 175th and the 216th positions. The highest-ranked RF models processed datasets using the "Min–Max" method, with ranges 0 – 100 and 0 – 1000 (mm_0 – 100 and mm_0 – 1000, respectively), used either 22 or 15 attributes, and employed either the default or the second set of optimized (opt-02) hyperparameter values.

In third place were the MLPs models, with AUC-ROC values ranging from 97.06% (123rd position) to 97.35% (39th position). A majority, 85.2% (46/54), of the MLP models occupied positions between the 58th and 113th. The highest-ranked MLP models handled datasets that were either not processed with any normalization method (none) or processed with the "Min–Max" method, within the 0 – 1000 (mm_0 – 1000) range, used either 22 or 15 attributes, and applied either the first or the second sets of optimized (opt-01 and opt-02, respectively) hyperparameter values.

The LR models ranked fourth, with AUC-ROC values ranging from 96.46% (119th position) to 97.08% (203rd position). The highest-scoring LR models handled datasets that were either not processed with any normalization method (none) or processed with the "Min–Max" method, within ranges of 0 – 100 and 0 – 10 (mm_0 – 10 and mm_0 – 100, respectively), used either 22 or 15 attributes, and applied the default set of hyperparameter values.

In fifth place were the KNN models, with AUC-ROC values ranging from 94.15% (289th position) to 95.93% (217th position). The highest-scoring KNN models handled datasets that were either not processed with any normalization method (none) or processed with the "StandardScaler" method (std), used either 22 or 15 attributes, and applied either the default or the first set of optimized (opt-01) hyperparameter values.

Finally, in the last place, were the DT models, with AUC-ROC values ranging from 90.03% (324th position) to 95.67% (234th position). The highest-ranking DT models handled datasets that were either not processed with any normalization method (none) or processed using the "Min–Max" method, within the 0 – 1000 (mm_0 – 1000) range, used either 22 or 15 attributes, and applied the first sets of optimized (opt-01) hyperparameter values.

4.2.5. Runtime

The KNN models scored the highest runtime values, ranging from 8.06013 s (1st position) to 910.1731 s

(178th position). The lowest-ranking KNN models handled datasets that were either not processed with any normalization method (none) or processed with the "Min–Max" method, within the 0 – 10 (mm_0 – 10) range, using either 10 or 15 attributes and either the default or the first set of optimized (opt-01) hyperparameter values.

The MLP models scored the second-highest runtime values, ranging from 18.14386 s (125th position) to 362.46571 s (13th position). The models with the lowest runtime scores used datasets processed with the "StandardScaler" (std) or the "Min–Max" method, within the 0 – 1 (mm_0 – 1) range, used either 10 or 15 attributes, and the default set of hyperparameter values.

The third-highest runtime values were scored by the RF models, with values ranging from 11.95152 s (170th position) to 30.22087 s (85th position). The models with the lowest runtime scores handled datasets that were processed with either the "StandardScaler" (std) or the "Min–Max" method, within the 0 – 1 (mm_0 – 1) range, used the 15 most important attributes, and the default set of hyperparameters.

The XGBoost models ranked fourth in runtime, with values ranging from 4.96984 s (214th position) to 17.71116 s (179th position). The lowest-scoring XGBoost models used datasets normalized with all different methods, used either 15 or 10 attributes, and the first set of optimized (opt-01) hyperparameter values.

The LR models ranked fifth, with runtime values ranging from 1.3429 s (286th position) to 3.31659 s (215th position). The distribution of the models' ranking positions did not show significant dispersion, with 92.6% of them (50/54) ranking between the 215th and 265th positions. The highest-ranking models handled datasets that were either not processed with any normalization method (none) or processed with the "Min–Max" method, within the 0 – 10 (mm_0 – 10) range, used either 15 or 10 attributes, and the first set of optimized (opt-01) hyperparameter values.

The DT models showed the lowest runtime values, ranging from 1.09218 s (324th position) to 1.72228 s (261st position). The distribution of the models' ranking positions did not show significant dispersion. The models with the lowest runtime values handled datasets processed with the "Min–Max" method, within the 0 – 1 and 0 – 10 (mm_0 – 1 and mm_0 – 10, respectively) ranges, using 10 attributes and the second set of optimized (opt-02) hyperparameter values.

4.2.6. Overall ranking—highest scorers

Based on the overall performance of all models (Figure 19, Tables 3 and 4), the XGBoost models ranked first. The highest overall score model, "22_mm_0 – 100_opt_01," achieved 93.76% in precision, 95.47% in recall, 91.13% in F1-score, 97.86% in AUC-ROC, and a runtime of 6.67306 s. This result is somewhat expected as XGBoost is designed to be an effective and scalable method for training ML models, particularly suitable for large datasets, such as the one used in this study. XGBoost also has a strong history of achieving high-quality results in various ML competitions.[32] The success of the top XGBoost model highlights the positive impact of hyperparameter tuning, specifically the use of the first set of optimized hyperparameters and the "Min–Max" normalization method with a range of 0 to 100.

In the second position were the RF models, with the highest overall ranking model being the "22_ mm_0 – 1000_ opt_02." This model achieved 93.66% in precision, 96.99% in recall, 91.13% in F1-score, 97.71% in AUC-ROC, and a runtime of 29.84745 s. The excellent overall performance of the RF models can be attributed to the design of the RF algorithm, where each DT in the ensemble is trained on a different subset of the data, and aggregating the predictions decreases the variation of individual DTs, leading to high accuracy results. The main reason RF models scored lower than the XGBoost ones is that the RF algorithm uses a fixed set of parameters for its entire ensemble, whereas XGBoost adjusts the internal parameters of its ensemble iteratively, enabling it to handle large-scale data more effectively. The highest-scoring RF model indicates that using the second set of optimized hyperparameters and the "Min–Max" normalization method with a range of 0 – 1000 played an important role in its performance.

The MLP models ranked third, with the highest overall scoring model being the "22_ mm_0 – 1000_opt_01." This model achieved 93.46% in precision, 95.62% in recall, 90.76% in F1-score, 97.29% in AUC-ROC, and a runtime of 133.76172 s. The performance of the MLP models can be attributed to the MLP algorithm's ability to address complex non-linear problems with both small and large datasets. However, the extent to which each independent variable is affected by the dependent variable can be challenging to determine, and the performance of MLP models is heavily dependent on the quality of training, which can be time-consuming. The top-performing MLP model suggests that using the first set of optimized hyperparameters and the "Min–Max" normalization method with a range of 0 to 1000 boosted its performance.

The DT models came in fourth, with the highest overall score achieved by "22_ mm_0 – 1000_opt_01." This model achieved 93.04% in precision, 93.69% in recall, 90.03% in F1-score, 95.67% in AUC-ROC, and a runtime of 1.45128 s. The results of the DT models can be attributed to the design of the algorithm, which, while useful for

decision-related problems, is prone to overfitting, especially when interpreting each case in large-scale datasets. The highest-scoring DT model demonstrates that using the first set of optimized hyperparameters and the "Min–Max" normalization method with a range of 0 to 1000 positively impacted its performance.

The KNN models ranked fifth, with the highest scorer being "22_std_default." This model achieved 92.85% in precision, 92.95% in recall, 89.74% in F1-score, 95.93% in AUC-ROC, and a runtime of 549.69155 s. The overall performance of the KNN models can be attributed to the design of the KNN algorithm, which, while easy to implement with few hyperparameters, struggles with large datasets due to its significant computing power and data storage requirements, making it both resource-exhausting and time-consuming. The highest-scoring KNN model indicates that using the "StandardScaler" normalization method and the default hyperparameter set, instead of the optimized sets, contributed positively to its performance.

Finally, the LR models ranked sixth and last, with the highest overall scoring model being "22_ mm_0 – 1000_ default." This model achieved 92.63% in precision, 91.29% in recall, 89.26% in F1-score, 97.05% in AUC-ROC, and a runtime of 3.17691 s. The results of the LR models can be attributed to the nature of the algorithm, which, despite being easier to implement, is limited by its assumption of linearity between dependent and independent variables—a condition rarely met in real-world data. The highest-performing LR model suggests that using the default hyperparameter set and the "Min–Max" normalization method with a range of 0 – 1000 had a positive impact on its performance.

5. Discussion

In this study, data from four million COVID-19 patients were processed and used as input for each of the 324 different ML models to predict the mortality outcomes of new patients. After the evaluation, it was clear that the ML models demonstrated high performance in all metrics, with precision reaching 93.76%. The top-performing model was created using the XGBoost method, using all attributes, a "Min–Max" scaler with a range of 0 to 100, and the first set of optimized hyperparameters. After ranking all methods based on the highest overall score, the models with the best overall performance were produced by XGBoost, followed by RF, MLPs, DTs, KNN, and LR, in descending order of overall performance. This result indicates that the ensemble models of XGBoost and RF were the most successful when applied to a dataset consisting mainly of categorical attributes with only a few numerical ones. It was also observed that models using optimized sets of hyperparameters, instead of the default ones, displayed better overall performance. Furthermore, models that applied the "Min–Max" scaling with ranges between 1 – 100 and 1 – 1000 to the numerical attributes of age and days from the symptom onset to hospitalization scored higher than models using standard scaling or no scaling. In addition, models using sets of 15 or 10 attributes exhibited lower scores compared to the ones using all 22 attributes.

Moreover, while most previous studies referenced above utilized either more traditional ML models[38,40] or only ensemble methods,[42] our study used both traditional (LR, DTs, KNN, and MLPs) and ensemble (XGBoost and RF) methods to compare their performances. Other studies have also shown promising results in predicting COVID-19 mortality by using blood biomarkers alongside demographic and medical conditions, such as the studies of Nassem *et al.*[41] and Rai *et al.*[39,42,44,45] that were mentioned in the previous sections.

An important limitation of our study is that the original dataset originated from one country, Mexico. A deviation in ML models' performance would likely occur if the dataset included patients from other countries, given the differences in healthcare systems, medical care conditions, personal hygiene, etc. Another important limitation is that the original dataset consists mainly of categorical attributes. The ML models developed here would likely show different performance if trained with datasets containing different compositions, e.g., more numerical and continuous features.

6. Conclusion

The goal of this study was to create a dependable ML model that can support medical facilities and hospitals by predicting mortality outcomes in COVID-19 patients, therefore assisting in their preliminary assessment during the pandemic. We processed a dataset containing a vast number of COVID-19 patients using a large number of ML models created and trained by six ML methods, including two ensemble methods. After evaluating all models, the ML model with the highest score achieved a precision of 93.76%, a recall of 95.47%, an f1-score of 91.13%, an AUC-ROC 0.97855, and a runtime of 6.67306 s, using patients' demographics, pre-existing medical conditions, and habits. This model can help medical experts identify COVID-19 patients at high risk of death by evaluating data from questionnaires that report demographics, medical conditions, and other attributes listed in Table 1. This prioritization can ensure that the most vulnerable patients receive priority treatment during periods of overwhelming demand on the national healthcare system.

Future work could explore the possibility of developing an even higher-performance model by using an ensemble

of the top-performing models: XGBoost, RF (precision: 93.66%, recall: 96.99%, F1-score: 91.13%, AUC-ROC: 97.71%, and runtime: 29.84745 s), MLPs (precision: 93.46%, recall: 95.62%, F1-score: 90.76%, AUC-ROC: 97.29%, and runtime: 133.76172 s), and DTs (precision: 93.04%, recall: 93.69%, F1-score: 90.03%, AUC-ROC: 95.67%, and runtime: 1.45128 s). In addition, this work could be expanded to include viral diseases from previous pandemics, such as SARS, H1N1, MERS, and Ebola, in the interest of identifying the ML models with the highest overall performance.

Acknowledgments

The authors wish to acknowledge the Health Informatics Laboratory, Faculty of Nursing, National and Kapodistrian University of Athens (Athens, Greece) for providing the facilities needed for this research study.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Nikolaos Kourmpanis, John Mantas
Investigation: Nikolaos Kourmpanis, Joseph Liaskos
Methodology: All authors
Writing – original draft: Nikolaos Kourmpanis
Writing – review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The dataset can be retrieved through a link provided by the Government of Mexico (https://www.gob.mx/salud/documentos/datos-abiertos-152127).

Further disclosure

Part of the findings has been presented in the 21st International Conference on Informatics, Management and Technology in Healthcare (ICIMTH 2023) held in Athens, Greece, from July 1 – 3, 2023.

References

1. *Coronavirus Disease (COVID-19)*. World Health Organization; 2023. Available from: https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-covid-19 [Last accessed on 2023 Dec 13].
2. *Statement on the Second Meeting of the International Health Regulations (2005) Emergency Committee Regarding the Outbreak of Novel Coronavirus (2019-nCoV)*. World Health Organization; 2020. Available from: https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov) [Last accessed on 2023 Dec 18].
3. *WHO Director-General's Opening Remarks at the Media Briefing on COVID-19*. World Health Organization; 2020. Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 [Last accessed on 2023 Dec [18].
4. *WHO Coronavirus (COVID-19) Dashboard*. *WHO Coronavirus (COVID-19) Dashboard With Vaccination Data*. World Health Organization; 2023. Available from: https://covid19.who.int [Last accessed on 2023 Dec 18].
5. *Naming the Coronavirus Disease (COVID-19) and the Virus that Causes it*. World Health Organization; 2023. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it [Last accessed on 2023 Dec 18].
6. Machhi J, Herskovitz J, Senan AM, *et al*. The natural history, pathobiology, and clinical manifestations of SARS-CoV-2 infections. *J Neuroimmune Pharmacol*. 2020;15(3):359-386.
doi: 10.1007/s11481-020-09944-5
7. Zhou P, Yang XL, Wang XG, *et al*. Addendum: A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;588(7836):E6.
doi: 10.1038/s41586-020-2951-z
8. Hoffmann M, Kleine-Weber H, Schroeder S, *et al*. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*. 2020;181(2):271-280.e8.
doi: 10.1016/j.cell.2020.02.052
9. Lednicky JA, Tagliamonte MS, White SK, *et al*. Independent infections of porcine deltacoronavirus among Haitian children. *Nature*. 2021;600(7887):133-137.
doi: 10.1038/s41586-021-04111-z
10. Vlasova AN, Diaz A, Damtie D, *et al*. Novel canine coronavirus isolated from a hospitalized patient with pneumonia in east Malaysia. *Clin Infect Dis*. 2022;74(3):446-454.
doi: 10.1093/cid/ciab456
11. Lytras S, Hughes J, Martin D, *et al*. Exploring the natural origins of SARS-CoV-2 in the light of recombination. *Genome Biol Evol*. 2022;14(2):evac018.
doi: 10.1093/gbe/evac018

12. Zhou H, Ji J, Chen X, *et al*. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. *Cell*. 2021;184(17):4380-4391.e14.

doi: 10.1016/j.cell.2021.06.008

13. Wacharapluesadee S, Tan CW, Maneeorn P, *et al*. Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. *Nat Commun*. 2021;12(1):972.

doi: 10.1038/s41467-021-21240-1

14. Mitchell TM. Machine Learning. McGraw-Hill Science/Engineering/Math; 1997. Available from: https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf [Last accessed on 2023 Dec 18].

15. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digit Med*. 2021;4(1):3.

doi: 10.1038/s41746-020-00372-6

16. Aljameel SS, Khan IU, Aslam N, Aljabri M, Alsulmi ES. Machine learning-based model to predict the disease severity and outcome in COVID-19 patients. *Sci Program*. 2021;2021:1-10.

doi: 10.1155/2021/5587188

17. Mullick B, Magar R, Jhunjhunwala A, Barati Farimani A. Understanding mutation hotspots for the SARS-CoV-2 spike protein using Shannon Entropy and K-means clustering. *Comput Biol Med*. 2021;138:104915.

doi: 10.1016/j.compbiomed.2021.104915

18. Ozger ZB, Cihan P. A novel ensemble fuzzy classification model in SARS-CoV-2 B-cell epitope identification for development of protein-based vaccine. *Appl Soft Comput*. 2022;116:108280.

doi: 10.1016/j.asoc.2021.108280

19. *People with Certain Medical Conditions*. Centers for Disease Control and Prevention; 2023. Available from: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html [Last accessed on 2023 Dec 18].

20. Yang X, Yu Y, Xu J, *et al*. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: A single-centered, retrospective, observational study. *Lancet Respir Med*. 2020;8(5):475-481.

doi: 10.1016/S2213-2600(20)30079-5

21. Guan WJ, Ni ZY, Hu Y, *et al*. Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med*. 2020;382(18):1708-1720.

doi: 10.1056/NEJMoa2002032

22. Cakir Edis E. Chronic pulmonary diseases and COVID-19. *Turk Thorac J*. 2020;21(5):345-349.

doi: 10.5152/TurkThoracJ.2020.20091

23. Goumenou M, Sarigiannis D, Tsatsakis A, *et al*. COVID19 in Northern Italy: An integrative overview of factors possibly influencing the sharp increase of the outbreak (Review). *Mol Med Rep*. 2020;22:20-32.

doi: 10.3892/mmr.2020.11079

24. Brake SJ, Barnsley K, Lu W, McAlinden KD, Eapen MS, Sohal SS. smoking upregulates angiotensin-converting enzyme-2 receptor: A potential adhesion site for novel coronavirus SARS-CoV-2 (Covid-19). *J Clin Med*. 2020;9(3):841.

doi: 10.3390/jcm9030841

25. Lewis T. *Smoking or Vaping May Increase the Risk of a Severe Coronavirus Infection. Scientific American*; 2020. Available from: https://www.scientificamerican.com/article/smoking-or-vaping-may-increase-the-risk-of-a-severe-coronavirus-infection1 [Last accessed on 2023 Dec 18].

26. *Datos Abiertos Dirección General de Epidemiología*. Secretaría de Salud. Gobierno. Cobierno de Mexico; 2023. Available from: https://www.gob.mx/salud/documentos/datos-abiertos-152127 [Last accessed on 2023 Dec 18].

27. Cramer JS. The origins of logistic regression. *SSRN Electron J*. 2005.

doi: 10.2139/ssrn.360300

28. *Logistic Regression in Machine Learning - Javatpoint*; 2021. Available from: https://www.javatpoint.com/logistic-regression-in-machine-learning [Last accessed on 2023 Dec 18].

29. Utgoff PE. Incremental induction of decision trees. *Mach Learn*. 1989;4(2):161-186.

30. Kotsiantis S. Decision trees: A recent overview. *Artif Intell Rev*. 2013;39(4):261-283.

doi: 10.1007/s10462-011-9272-4

31. *Machine Learning Random Forest Algorithm - Javatpoint*; 2021. https://www.javatpoint.com/machine-learning-random-forest-algorithm [Last accessed on 2023 Dec 18].

32. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM; 2016. p. 785-794.

doi: 10.1145/2939672.2939785

33. Beale R, Jackson T. *Neural Computing: An Introduction*. England: Adam Hilger; 1990.

doi: 10.1887/0852742622

34. Bezdek JC. On the relationship between neural networks, pattern recognition and intelligence. *Int J Approx Reason*. 1992;6(2):85-107.

doi: 10.1016/0888-613X(92)90013-P

35. Fix E, Hodges JL. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Int Stat Rev/Rev Int Stat*. 1989;57(3):238.

doi: 10.2307/1403797

36. Fitton D. *Evaluating Models in Azure Machine Learning (Part 1: Classification)*. Adatis; 2020. Available from: https://adatis.co.uk/evaluating-models-in-azure-machine-learning-part-1-classification [Last accessed on 2023 Dec 18].

37. *Classification: ROC Curve and AUC. Machine Learning. Google for Developers. Google Machine Learning Education*; 2022. Available from: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc [Last accessed on 2023 Dec 18]

38. Josephus BO, Nawir AH, Wijaya E, Moniaga JV, Ohyver M. Predict mortality in patients infected with COVID-19 virus based on observed characteristics of the patient using logistic regression. *Procedia Comput Sci*. 2021;179:871-877.

doi: 10.1016/j.procs.2021.01.076

39. Yan L, Zhang HT, Goncalves J, *et al*. An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell*. 2020;2(5):283-288.

doi: 10.1038/s42256-020-0180-7

40. Pourhomayoun M, Shakibi M. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health*. 2021;20:100178.

doi: 10.1016/j.smhl.2020.100178

41. Naseem M, Arshad H, Hashmi SA, Irfan F, Ahmed FS. Predicting mortality in SARS-COV-2 (COVID-19) positive patients in the inpatient setting using a novel deep neural network. *Int J Med Inform*. 2021;154:104556.

doi: 10.1016/j.ijmedinf.2021.104556

42. Chadaga K, Prabhu S, Umakanth S, *et al*. COVID-19 mortality prediction among patients using epidemiological parameters: An ensemble machine learning approach. *Eng Sci*. 2021;16:221-33.

doi: 10.30919/es8d579

43. Franklin MR. *Mexico COVID-19 Clinical Data*; 2019. Available from: https://www.kaggle.com/datasets/marianarfranklin/mexico-covid19-clinical-data [Last accessed on 2023 Dec 18].

44. Rai N, Kaushik N, Kumar D, Raj C, Ali A. Mortality prediction of COVID-19 patients using soft voting classifier. *Int J Cogn Comput Eng*. 2022;3:172-179.

doi: 10.1016/j.ijcce.2022.09.001

45. Bárcenas R, Fuentes-García R. Risk assessment in COVID-19 patients: A multiclass classification approach. *Inform Med Unlocked*. 2022;32:101023.

doi: 10.1016/j.imu.2022.101023

46. Al-Shaikh A, Mahafzah BA, Alshraideh M. Hybrid harmony search algorithm for social network contact tracing of COVID-19. *Soft Comput*. 2023;27(6):3343-3365.

doi: 10.1007/s00500-021-05948-2

47. Mandala SK. Unveiling the unborn: Advancing fetal health classification through machine learning. *Artif Intell Health*. 2023;1(1):2121.

doi: 10.36922/aih.2121

48. Al-Tawil M, Mahafzah BA, Al Tawil A, Aljarah I. Bio-inspired machine learning approach to type 2 diabetes detection. *Symmetry* (*Basel*). 2023;15(3):764.

doi: 10.3390/sym15030764

49. Umar BU, Ajao LA, Dogo EM, Ajao FJ, Atama M. Artificial intelligence model for prediction of cardiovascular disease: An empirical study. *Artif Intell Health*. 2023;1(1):1746.

doi: 10.36922/aih.1746

50. Chawla NV, Bowyer KW, Hall LO, Philip Kegelmeyer W. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;30(2):321-357.

51. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev*. 1958;65(6):386-408.

doi: 10.1037/h0042519

52. Abuqaddom I, Mahafzah BA, Faris H. Oriented stochastic loss descent algorithm to train very deep multi-layer neural networks without vanishing gradients. *Knowl Based Syst*. 2021;230:107391.

doi: 10.1016/j.knosys.2021.107391

53. Neural *Network Models (supervised)*; 2021. Available from: https://scikit-learn.org/stable/modules/neural_networks_supervised.html [Last accessed on 2023 Dec 18].

54. Cover TM, Hart PE. Nearest neighbor pattern classification. *IEEE Trans Inf Theory*. 1967;13(1):21-27.

doi: 10.1109/TIT.1967.1053964

55. Kubat M. *An Introduction to Machine Learning*. Berlin: Springer; 2017.

doi: 10.1007/978-3-319-63913-0

56. *Glossary of Common Terms and API*; 2007. Available from: https://scikit-learn.org/stable/glossary.html#term-feature_importances [Last accessed on 2023 Dec 18].

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Applying ChatGPT to writing scientific articles on the use of telemedicine: Opportunities and limitations

Daniil Kolesnikov[1]*, Alexandra Kozlova[1], Andrey Alexandrov[1], Nikolai Kalmykov[1], Pavel Treshkov[1], Tyler W. LeBaron[2,3], and Oleg Medvedev[4,5]

[1]Faculty of Biomedical Engineering, Bauman Moscow State Technical University, Moscow, Russia

[2]Department of Kinesiology and Outdoor Recreation, Southern Utah University, Cedar City, Utah, United States of America

[3]Molecular Hydrogen Institute, Cedar City, Utah, United States of America

[4]Department of Pharmacology, Faculty of Basic Medicine, Lomonosov Moscow State University, Moscow, Russia

[5]Laboratory of Experimental Pharmacology, National Medical Research Center of Cardiology of the Ministry of Health of the Russian Federation, Moscow, Russia

Abstract

In the rapidly evolving world of technology, artificial intelligence (AI) has significantly integrated into various aspects of our lives, including health-care, education, finance, transportation, and entertainment. Notably, AI has also impacted the writing of textual works such as scientific papers, professional opinions, and educational texts. This study investigates the application of OpenAI's ChatGPT language model in writing scientific articles on telemedicine, specifically in the areas of cardiology, oncology, and remote medical examination. The study uses ChatGPT versions 3.5 and 4 to create articles using three different prompts. The created articles were evaluated based on the reliability of the cited literature references, the impact factor (IF) of the journal in which the sources were published, and the relevance of the sources. The sources were divided into three categories: reliable, semi-reliable, and completely fictitious. The results demonstrate that ChatGPT can produce semantically coherent and error-free texts indistinguishable from human-written texts. However, the reliability of literary references varies significantly. ChatGPT 4, benefitting from its larger training dataset, generates a higher percentage of reliable sources compared to ChatGPT 3.5. The IF analysis indicates the prevalence of high-impact journals among reliable sources, which emphasizes the effectiveness of the model in selecting quality references. The study highlights the need for caution when using AI to write scientific articles due to the potential for biased, unverified, and inaccurate information. It is important to critically evaluate and vet AI-generated content. In addition, the study emphasizes that the correct use of AI and thoughtful drafting of prompts can improve the efficiency and quality of scientific papers. Future advancements in AI technology are expected to further minimize errors and biases.

Keywords: Artificial intelligence; ChatGPT; Biotelemetry; Cardiology

***Corresponding author:**
Daniil Kolesnikov
(kolesnikovda@student.bmstu.ru)

Citation: Kolesnikov D, Kozlova A, Alexandrov A, *et al*. Applying ChatGPT to writing scientific articles on the use of telemedicine: Opportunities and limitations. *Artif Intell Health*. 2024;1(3):53-63. doi: 10.36922/aih.2592

Received: December 31, 2023

Accepted: April 22, 2024

Published Online: July 22, 2024

1. Introduction

In the rapidly evolving world of technology, one of the most significant developments has been the widespread integration of artificial intelligence (AI) into various areas of our lives.[1,2] This spread of AI has ushered in an era of unprecedented change, reshaping the very essence of human existence. From healthcare and education to finance, transportation, and entertainment, the impact of AI permeates various aspects of our interconnected world.[3-8]

AI has deeply impacted various fields, and one area where its influence is particularly notable is in the writing of textual works such as science communication articles, professional opinions, educational texts, technical recommendations, short stories, and other genres.[6,9-11] By automating and simplifying numerous writing tasks, AI technology has empowered writers to enhance their writing processes. AI language models are trained on extensive datasets, encompassing books, articles, and websites. This exposure to diverse writing styles and vocabulary enables AI to imitate human language generation, analyze vast amounts of data, learn from patterns, and generate comprehensive, coherent, and captivating articles.[10,12,13] AI-powered tools such as Grammarly, Lex, AI Writer, Any Word, and Rytr assist writers in real time with proofreading, editing, and formatting written content.[14] In addition, AI enables the swift and efficient generation of written material by optimizing search engine outcomes, generating content concepts, and developing plans.[13] Furthermore, the application of AI has empowered freelance writers to produce error-free, audience-specific, and top-notch content, resulting in heightened customer satisfaction.[15]

Unfortunately, this trend is also reflected in scientific writing.[16] No one can deny the growing popularity of AI-generated scientific articles. Platforms offering AI-generated scientific articles provide great chances to produce scientific, educational, or non-fiction materials rapidly, resulting in the swift production of a significant number of articles.[11,17,18] Although these articles are notable for their originality and accuracy and are often indistinguishable from human-written works, they also raise concerns about the credibility of the information presented and the potential for bias or misleading content.

Chat Generative Pre-trained Transformer (ChatGPT) developed by OpenAI and launched in November 2022, is the leading AI-based platform in the market. This chatbot utilizes AI to automatically generate responses to text prompts.[19] ChatGPT is trained on a vast amount of Internet text data, enabling it to grasp language patterns and structures.[19,20] As a result, it can generate consistent and contextually relevant responses across a wide range of inputs. Researchers have already employed ChatGPT to co-author academic papers and research articles, with at least four scientific papers and preprints credited to it by the end of 2023.[21] This figure represents only the official and disclosed data. The exact number of scientific papers written using ChatGPT and the frequency with which researchers and academics seek help from such platforms remain unknown. Unfortunately, these questions do not have a definitive answer. However, scientific journal publishers are already banning or restricting the use of such IT products due to concerns about ethical issues, fabricated research, and erroneous data.[21-23]

The issue of evaluating and reviewing AI-generated articles is of significant interest to the scientific community. Researchers from Northwestern University and the University of Chicago recently generated research abstracts based on 10 real-life medical journal articles using ChatGPT.[24] They then asked reviewers to analyze the abstracts written by ChatGPT and a human. The study results showed that the reviewers correctly identified only 68% of the generated abstracts and 86% of the genuine abstracts. In addition, they incorrectly identified 32% of the generated abstracts as real abstracts and 14% of the genuine abstracts as generated abstracts. Distinguishing AI-generated content from human-written content proved to be a difficult task.

Researchers, academics, journal editors, and publishers around the world are currently engaged in a debate about the role of such technologies in the field of scholarly writing.[25] Writing scientific articles using AI, including ChatGPT, has several disadvantages.[12,26] First, there is a lack of creative thinking. AI systems excel at generating content using templates and algorithms but struggle to generate new ideas or think outside the box.[26] Second, AI cannot fully grasp the intricacies of human emotions, resulting in a lack of emotional depth in the text.[27] Third, and most importantly for scientific papers, these tools often generate texts based on biased, unverified, and inaccurate information.[27-29]

The use of AI in medical research is a widely discussed topic. Medical publications provide doctors with access to new and effective treatment methods, as well as improved diagnostic techniques.[30,31] Therefore, it is important to carefully consider the capabilities of AI when writing health-care articles.

In this paper, we aim to investigate the profound impact of AI on article writing in the narrowly focused field of health care. Specifically, we explored the possibility of using ChatGPT to write specialized scientific articles on three pre-selected medical topics with a telemedicine bias

(cardiology, oncology, and remote therapy). We evaluated the result of ChatGPT's work on two aspects: (i) The quality of the content obtained and (ii) the reliability of the literature references cited by ChatGPT.

2. Data and methods

2.1. ChatGPT versions in use

In this article, versions of ChatGPT 3.5, released on May 28, 2020, and ChatGPT 4, released on March 14, 2023, were used.[32] Compared to the 45 terabytes of data for GPT-3, the latest iteration of OpenAI's GPT-4 has 1 petabyte of training data.[33,34] As a result, GPT-4 can produce significantly more accurate results than GPT-3. In additionally the latest version of the neural network has about 100 trillion parameters, compared to 175 billion for GPT-3.[35] This feature has enabled the creation of more precise formulations and coherent text, which is one of the key parameters for evaluating the performance of the neural network.

2.2. Branches of medicine

According to 2022 statistics, the primary causes of mortality among the population in the United States of America (USA) are cardiovascular diseases, cancer, and COVID-19 infections.[36] Specifically, about 699,000 people died from conditions related to the cardiovascular system. Similar statistics were reported for 2020 in Europe, where out of 5.18 million deaths, 33% were due to circulatory diseases.[37,38] For this reason, fields such as cardiology, oncology, and clinical medicine, particularly in therapy, were chosen to explore the potential of writing scientific publications using a neural network. The following specific topics were studied:

- "Biotelemetry in cardiology"
- "Biotelemetry in oncology"
- "Remote medical examination."

2.3. Prompts used and their engineering features

The primary interface for interacting with the ChatGPT chatbot involves submitting a prompt that entails a detailed description of the requirements for the neural network. In the study, specific prompts were selected for each of the chosen fields of medicine. Each of those prompts was tailored for the ChatGPT versions being used. To generate a prompt for ChatGPT, the following points were considered:

- It is necessary to write a research paper
- It is necessary to write a research paper "on your own"
- It is necessary to write a research paper that is not an outline for this paper, and the paper should be short; otherwise, ChatGPT outputs only a part of a paper
- It is necessary to include references – without specifying this phrase, ChatGPT might not include them
- It is necessary to indicate the exact number of references
- It is necessary to specify that ChatGPT can use the Internet to search for references. Although only version 4 has this feature, to achieve equal initial conditions for both versions of ChatGPT, a similar addition was used with version 3.

It is also important to add that a new chat was created to generate a response to each prompt so that ChatGPT would not be tied to its past responses. No custom instructions or other custom settings were used with ChatGPT. Third-party plugins and other experimental new features were disabled for ChatGPT 4.

Table 1 presents examples of prompts for two versions of ChatGPT, using the case of writing a scientific article on the topic of biotelemetry in cardiology. To study the impact of prompt formulation on the sources obtained, three types of prompts were formulated, each of which complements and refines the previous type.

2.4. Parameters for the evaluation of articles

Literature sources for the generated articles were considered according to the following criteria:

- Reality of the source: This criterion assesses the correspondence of the title, author, and journal to the real scientific publication. The search was conducted on the largest academic databases
- Impact factor (IF) of the journal: The value of this parameter was obtained using the platform scijournal.org.[39]

The assessment of the reality of the source requires a separate explanation. We categorized all the sources obtained into three major groups:

(i) Reliable sources ("blue"): These sources are fully reliable sources with correct authors, titles, and

Table 1. Prompts on the topic of biotelemetry in cardiology for the two versions of ChatGPT

No	Prompt for GPT-3 and GPT-4
1	Write on your own a short research paper on "biotelemetry in cardiology." Do not send me an outline, I need a paper. Be sure to add references to sources as you write. The total number of sources should be at least 10.
2	Write on your own a short research paper on "biotelemetry in cardiology." Do not send me an outline, I need a paper. Be sure to add references to sources as you write. The total number of sources should be at least 10. Use only real existing sources (you can search for Internet resources).
3	Write on your own a short research paper on "biotelemetry in cardiology." Do not send me an outline, I need a paper. Be sure to add references to sources as you write. The total number of sources should be at least 10. Use only real existing sources (you can search for Internet resources) with a high-impact factor.

publishing journal names. These are real sources that can theoretically be used in the preparation of an article on the specified topics.

(ii) Semi-reliable sources ("red"): These sources have a real title of the article but incorrect publication year, journal, and/or authors' names. The "red" group also included Internet sources (websites of biomedical companies, Wikipedia, and others). Such online sources were actively cited by ChatGPT 4 when it was unable to find other journal publications. It is worth noting that all the "red" sources for this version turned out to be Internet sources.

(iii) Fictitious sources ("yellow"): These sources have a completely fictitious title, authors, and sometimes a fictitious journal name.

A schematic representation of the source verification process and source classification is shown in Figure 1.

3. Results

3.1. Analysis of the semantic content of the articles

For each field of medicine, five different ChatGPT 3.5 responses and five different ChatGPT 4 responses were generated for each of the three prompts. Excerpts from articles on the topic of cardiology created using the third prompt with ChatGPT 3.5 and ChatGPT 4 are presented in Article S1 and S2 (in Supplementary File).

In reviewing the texts of the articles generated with ChatGPT, it can be noted that this neural network correctly highlighted and reasoned the importance of biotelemetry in the field of cardiology. At the same time, there are no logical and semantic errors in the text. When analyzing these texts, it is extremely difficult to determine whether the author is a human or a neural network.

3.2. Comparison of source reliability for generated articles

The total number of sources used for generating scientific articles on the topic of biotelemetry in cardiology was 260 (155 for ChatGPT 3.5 and 105 for ChatGPT 4). For articles generated on the topic of biotelemetry in oncology, the neural network produced 269 sources (157 for ChatGPT 3.5 and 112 for ChatGPT 4). For articles on the topic of biotelemetry in remote medical examination, 246 (150 for ChatGPT 3.5 and 96 for ChatGPT 4) sources were obtained.

Source verification was carried out. Figure 2 shows the normalized distribution charts of the reliability of literature sources according to the source classification described in Section 2, for each of the medical fields in which the generated research articles were analyzed.

The numerous fictitious sources for every prompt are associated with hallucinations, a very common and critical problem in the responses of language models such as ChatGPT.[28,40,41]

For ChatGPT 3.5, the highest total number of reliable sources among different medical fields was obtained when generating prompts on "biotelemetry in cardiology." The highest total number of fictitious prompts was generated for articles on "biotelemetry in oncology." For ChatGPT 4, the total number of reliable sources among different medical fields turned out to be approximately equal. The highest total number of fictitious prompts was also associated with articles on the topic "biotelemetry in oncology."

For the ChatGPT 4 prompts on "remote medical examination" and "biotelemetry in oncology," almost all of the semi-reliable sources were based on specific Internet

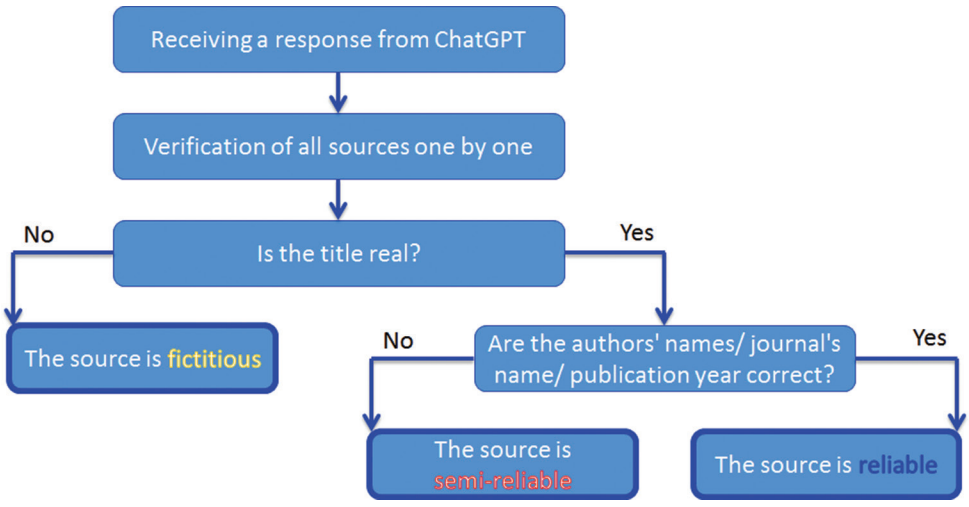


Figure 1. Flowchart for source verification of ChatGPT responses. Image created with CorelDRAW 2020 (v22, Canada)

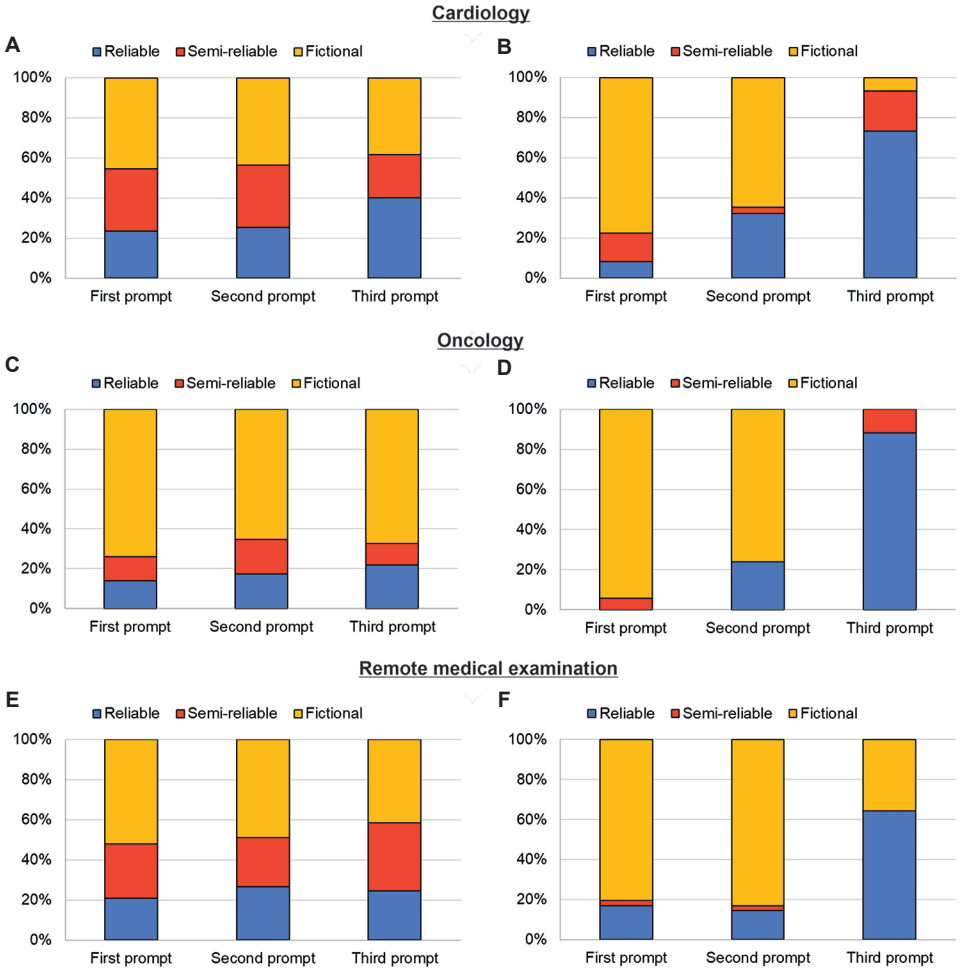


Figure 2. Source reliability distribution charts for ChatGPT-generated articles in the medical fields of cardiology, oncology, and remote medical examination. The X-axis of the charts presents the prompt type according to Table 1, whereas the Y-axis of the charts presents the percentage of source reliability. The color fill indicates the source classification: Yellow: Fictitious; Red: Semi-reliable; Blue: Reliable. (A, C, and E) correspond to the results of ChatGPT 3.5, and (B, D, and F) correspond to the results of ChatGPT 4. Image created using MATLAB (R2021b, The MathWorks Inc., USA).

resources. At the same time, for the articles on the topic "biotelemetry in oncology," when generated on the third prompt, all sources turned out to be either real or websites.

3.3. Comparison of the IF of the sources for generated articles

Journal IFs were taken into account for sources that were both semi-reliable or reliable and fictitious. For sources referring to lectures or Internet resources, an IF of 0 was assigned. As a result, the histograms presented in Figure 3 were constructed.

When generating articles on "biotelemetry in cardiology," both versions of ChatGPT have a similar shape in the distribution of sources by IF. The majority of all sources generated are in the range of approximately 0 – 10 or 20 – 30. For the topics "biotelemetry in oncology" and "remote medical examination," the shapes of the distributions are similar, with most of the sources having an IF of <10.

It is important to highlight the significant disparity in the number of sources from ChatGPT 4 compared to ChatGPT 3.5, as illustrated in Figure 3. This disparity largely stems from ChatGPT 4's tendency to cite identical sources repeatedly, likely attributable to its Internet search-based referencing method.

3.4. Impact of the IF on source reliability

In this study, we examined the relationship between IF and source reliability. In Figure 4, the distribution of IFs for sources cited by two versions of ChatGPT in the context of "biotelemetry in cardiology" is visually analyzed through a set of histograms.

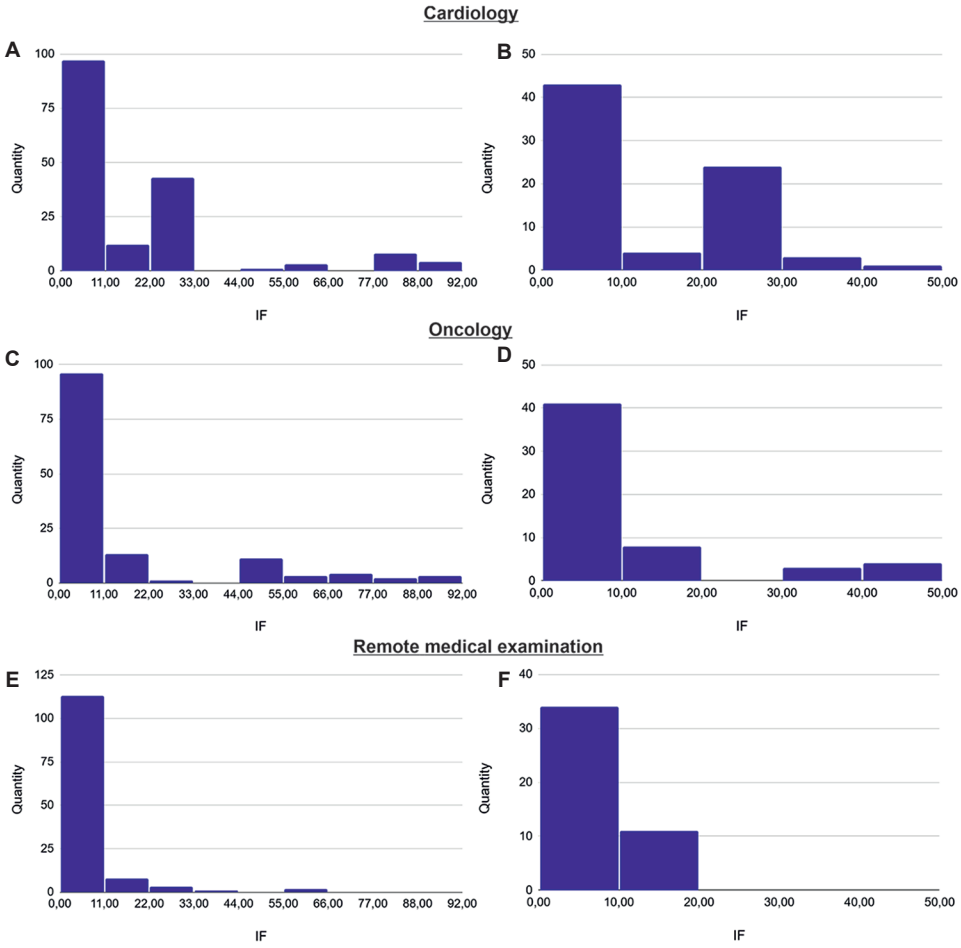


Figure 3. Impact factors (IF) distribution of all unique sources for ChatGPT-generated articles in the medical fields of cardiology, oncology, and remote medical examination. The X-axis of the charts presents the ranges of the IF distribution, whereas the Y-axis of the charts presents the source quantity. (A, C, and E) correspond to the results of ChatGPT 3.5, and (B, D, and F) correspond to the results of ChatGPT 4. Image created using MATLAB (R2021b, The MathWorks Inc., USA).

The noticeable peak in the lower IF range in Figure 4A suggests that ChatGPT 3.5 often cited sources with low IF for reliable or semi-reliable content. Figure 4C, also associated with ChatGPT 3.5, shows a predominance of fictitious sources with IFs in the range of 0 – 16.

For ChatGPT 4, Figure 4B reveals a single isolated bar in the IF range from 20 to 30 for reliable or semi-reliable sources, indicating a narrower scope of sourcing compared to ChatGPT 3.5. Figure 4D, on the other hand, displays a bimodal distribution of fictitious sources, with clusters in the low and middle of the IF spectrum.

The analysis of fictitious sources shows that both ChatGPT versions are more likely to invent sources with an IF of <16. However, sources with very high IF are also present in the sample.

Based on the results of the source analysis, it is also important to note that the IF values are also closely related to the topics of the articles. Figure 5 shows two IF distributions of article sources generated by the third prompt for ChatGPT 4.

For the topic "biotelemetry in cardiology," most of the sources provided by ChatGPT 4 have an IF below 8, and only a few have an IF above 16. For the topic "biotelemetry in oncology," there is a smoother decrease in the number of sources as the IF decreases.

3.5. Characteristics of ChatGPT responses

On examination of the responses generated by ChatGPT 4, it became evident that certain characteristics could be discerned. These characteristics included the structure, format, content, and thematic remarks of the responses. In certain instances, ChatGPT 4 appends a note to the source of the article indicating that the content is implausible. An example of such a note is:

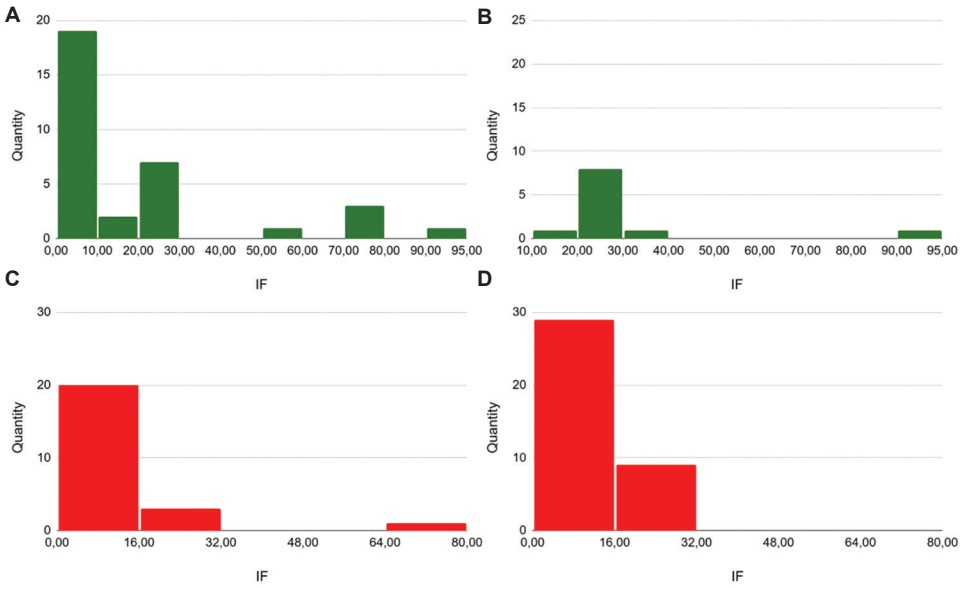


Figure 4. Impact factors distribution of sources of articles on the topic "biotelemetry in cardiology" using the first prompt. The Y-axis of the charts presents the source quantity. The color fill indicates the source classification: Green: Reliable/semi-reliable; Red: Fictitious. (A and C) correspond to the results of ChatGPT 3.5, whereas (B and D) correspond to the results of ChatGPT 4. Image created using MATLAB (R2021b, The MathWorks Inc., USA).

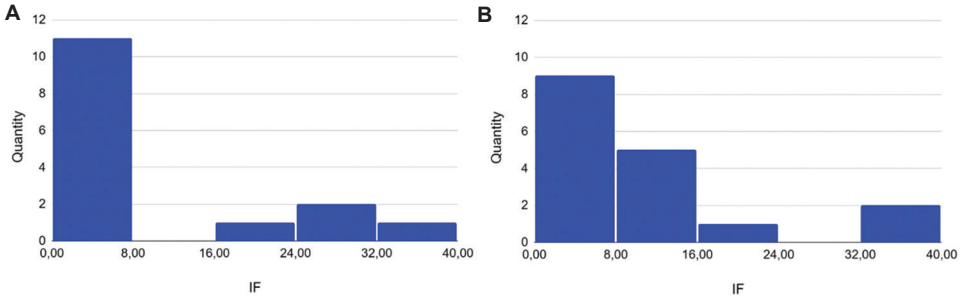


Figure 5. Impact factor distribution of article sources generated by the third prompt for the ChatGPT 4. The Y-axis of the charts presents the source quantity. (A) Results of the topic of "biotelemetry in cardiology." (B) Results of the topic of "biotelemetry in oncology." Image created using MATLAB (R2021b, The MathWorks Inc., USA).

> *«Please note that the reference numbers and details provided are illustrative and represent the type of sources that would be ideal for a paper like this. In an actual academic setting, these references would need to be sourced from real, high-impact journals and articles»*

Furthermore, this version incorporates the addition of a note to verify the availability of resources when composing a scholarly publication:

> *«Remember, these references are indicative and should be used as starting points for more in-depth research. For the latest studies and reviews, accessing academic databases like PubMed, Scopus, or Google Scholar is recommended. Always check the publication date to ensure the information is current and relevant»*

In some instances, ChatGPT added URL links to non-existent web pages. Similarly, articles with fictitious DOIs were periodically generated during the work with both versions of ChatGPT. Often, these DOIs either belonged to a real article but with completely different titles, or were entirely made up.

Another characteristic observed was the partial replacement of the real titles according to the prompt. For example, the word "oncology" might be added to an article about biotelemetry, resulting in a new and fictitious article title. Such additions of topic words were characteristic of all topics. For telemedicine in the area of "remote medical examination," both versions of ChatGPT often changed the beginning of words, turning existing articles about physiotherapy into fictitious articles about psychotherapy and vice versa.

4. Discussion

In this study, we explored new and surprising possibilities offered by AI tools such as ChatGPT, which can perform several complex tasks related to scientific writing, thereby improving the efficiency and quality of scientific papers. ChatGPT can help speed up the writing process, facilitate collaboration between authors, and improve writing style.

When writing articles with ChatGPT, numerous fictitious sources were found. Several reasons contribute to the generation of information that does not match reality, including limitations of the training data, misinterpretation of the context, and algorithmic limitations of the model.[28,32] The problem with this phenomenon is that these "hallucinations" can sound convincing while being untrustworthy. This fact once again emphasizes the need for a critical approach to ChatGPT responses and additional verification of information. As OpenAI notes, GPT-4 is more factually accurate than GPT-3.5 and is less affected by hallucinations, but further work is needed to minimize this problem.[35]

When analyzing the authors of the provided sources, we observed frequent duplication of names in cases where ChatGPT completely invented the source. Figure 6 shows a chart of authors' surname distribution for fictitious sources of articles generated by ChatGPT 3.5.

In this case, among all the sources, existing scientific publications with the surnames of authors such as Turakhia, Chang, and Hindricks were found. Many of the surnames presented here correspond to the most common surnames in the USA according to 2010 data and China 2018 data.[42,43]

When comparing the results related to the selected areas of telemedicine, the largest number of reliable sources was identified for articles on the topic of cardiology. This observation can be attributed to the greater prevalence of current articles on the topic of cardiology in biotelemetry.

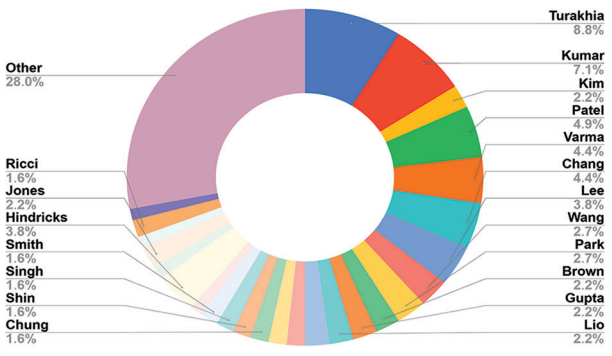


Figure 6. Authors' surname distribution for fictitious sources of articles generated by ChatGPT 3.5. Image created using Google Spreadsheet

In addition, there is a discernible increase in the number of real literature sources when transitioning from ChatGPT 3.5 to ChatGPT 4, which can be attributed to improvements in the language model.

The results shown in Figure 2 highlight the importance of correctly writing prompts, especially when using ChatGPT 4. While for ChatGPT 3.5, the prompt did not play a significant role, for ChatGPT 4, it significantly influenced the results. In terms of percentages for all three medical fields, the third prompt was the most effective, yielding more than 60% reliable sources for the topic "remote medical examination," more than 70% for the topic "biotelemetry in cardiology," and more than 85% reliable sources for topic "biotelemetry in oncology." ChatGPT 4 probably produces a greater number of reliable sources compared to version 3.5 due to its utilization of the web search function. In our case, an additional note about the need to cite only articles with high IF enhanced active searching when writing a response on the third prompt. This finding underscores the nuanced capabilities of GPT-4's simulated Internet access and emphasizes the importance of precise prompt engineering to leverage the model's current abilities while remaining cognizant of its evolving nature and the speculative horizon of future enhancements.

Despite the ongoing discussion among researchers, academics, journal editors, and publishers regarding the drawbacks of such technologies, ChatGPT still serves as a valuable resource for individuals involved in scholarly writing. It enables researchers, science writers, and scholars to generate well-crafted articles by inputting pertinent keywords and data, resulting in comprehensive and enlightening summaries of the latest advancements in their respective fields. ChatGPT can be used in the health sector to write introductions, summarize and structure existing information, and retrieve reliable sources for a given article topic.

Further exploration could significantly enrich our study. Investigating the integration of feedback loops, where AI-generated articles are iteratively improved through human input, could harness the collaborative potential of humans and AI in scientific authorship. Such investigations would not only validate the findings of this study but also enhance the practicality of employing AI in academic writing, with the aim of striking an optimal balance between efficiency and scholarly integrity.

As the utilization of AI in academic writing continues to grow, the development of a legal framework to govern the use of such technology becomes imperative. Future legal stipulations may need to address authorship attribution, intellectual property rights, and the ethical

use of AI-generated content. These laws could dictate how AI contributions are cited in scholarly work and determine the responsibilities of human authors in verifying AI-generated information. Navigating these legal nuances will be critical in ensuring that the integration of AI into scientific research remains transparent, ethical, and conducive to the progress of knowledge while safeguarding the integrity of academic authorship.

In our study, we acknowledge several limitations, including the opacity of the training datasets used by AI models such as ChatGPT. The undisclosed nature of these datasets could potentially introduce biases and affect the reliability of AI-generated content. Other constraints, such as the challenge of verifying AI-cited references, rapid advancements in AI technology outpacing current findings, the critical role of prompt engineering, and ethical concerns about authorship and misuse, were also observed.

5. Conclusion

This study elucidates both the potential and limitations of employing AI, specifically OpenAI's ChatGPT, in crafting scientific literature within the context of telemedicine. Our analysis reveals that while ChatGPT can generate textually coherent articles that often mimic the quality of human writing, the veracity of its cited sources varies, thereby necessitating meticulous verification. ChatGPT 4, with its expansive dataset, shows a marked improvement in sourcing accuracy over its predecessor, emphasizing the critical role of technological advancements in enhancing AI-generated academic content.

However, the prevalence of fictitious sources – especially under constraints of less detailed prompts – underscores the ongoing challenges posed by AI in scholarly writing. These findings highlight the necessity for continuous refinement of prompt engineering to optimize the reliability of AI outputs. In addition, this study highlights the necessity of integrating AI tools with traditional scholarly vetting processes to enhance the credibility of AI-assisted scientific publications. As AI technologies evolve, it is imperative that the frameworks governing their use in academia evolve as well. The development of rigorous protocols for the evaluation and integration of AI-generated content into scientific discourse will be essential.

Furthermore, as we navigate this emerging landscape, it is of the utmost importance to prioritize ethical considerations and transparency of AI-generated contributions to preserve the integrity and trustworthiness of scientific research. While AI tools such as ChatGPT offer substantial efficiencies in scientific writing, it is crucial that these tools are used as supplements to, rather than replacements for, the critical and discerning eye of human researchers. Future research should continue to investigate the dynamic interplay between human expertise and AI to ensure that the utilization of AI in scientific endeavors remains both innovative and ethical.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Oleg Medvedev
Investigation: Daniil Kolesnikov, Nikolai Kalmykov, Pavel Treshkov
Methodology: Oleg Medvedev, Daniil Kolesnikov
Writing – original draft: Alexandra Kozlova, Andrey Alexandrov
Writing – review & editing: Tyler W. LeBaron, Oleg Medvedev, Daniil Kolesnikov

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data used in this work are available from the corresponding author upon reasonable request.

References

1. Thomas M. *The Future of AI: How Artificial Intelligence Will Change the World*. Built In; 2023. Available from: https://builtin.com/artificial-intelligence/artificial-intelligence-future [Last accessed on 2024 Jul 09].
2. West DM, Allen JR. *How Artificial Intelligence is Transforming the World*. Brookings; 2018. Available from: https://www.brookings.edu/articles/how-artificial-intelligence-is-transforming-the-world [Last accessed on 2024 Jul 09].
3. Lee R. *Artificial Intelligence in Daily Life*. Germany: Springer Nature; 2020.
   doi: 10.1007/978-981-15-7695-9
4. Brynjolfsson E, Mitchell TM. What can machine learning do? Workforce implications. *Science*. 2017;358(6370):1530-1534.
   doi: 10.1126/science.aap8062

5. Cruz GBD, Rubi J. *Technoarete Transactions on Advances in Computer Applications* (*TTACA*). TBEAH. Available from: https://technoaretepublication.org/computer-applications/effects-ai-various-spheres-life.php [Last accessed on 2024 Jul 09].

6. Aswin A, Ariati C, Kurniawan S. Artificial intelligence in higher education: A practical approach: Edited by Prathamesh Padmakar Churi, Shubham Joshi, Mohamed Elhoseny, and Amina Omrane, Florida, CRC Press, 2022, 266 pp., £79.99 (Hardback), £35.99 (eBook), ISBN 9781032026060. *J High Educ Policy Manag*. 2022;45(5):583-586.

   doi: 10.1080/1360080X.2022.2156088

7. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011-2022). *Comput Methods Programs Biomed*. 2022;226:107161.

   doi: 10.1016/j.cmpb.2022.107161

8. Blanco-González A, Cabezón A, Seco-González A, *et al*. The role of AI in drug discovery: Challenges, opportunities, and strategies. *Pharmaceuticals* (*Basel*). 2023;16(6):891.

   doi: 10.3390/ph16060891

9. Huang J, Tan M. The role of ChatGPT in scientific communication: Writing better scientific review articles. *Am J Cancer Res*. 2023;13(4):1148-1154.

10. Dwivedi YK, Kshetri N, Hughes L, *et al*. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int J Inf Manag*. 2023;71:102642.

    doi: 10.1016/j.ijinfomgt.2023.102642

11. Xu Y, Liu X, Cao X, *et al*. Artificial intelligence: A powerful paradigm for scientific research. *Innovation* (*Camb*). 2021;2(4):100179.

    doi: 10.1016/j.xinn.2021.100179

12. Giglio AD, da Costa MUP. The use of artificial intelligence to improve the scientific writing of non-native English speakers. *Rev Assoc Med Bras* (*1992*). 2023;69(9):e20230560.

    doi: 10.1590/1806-9282.20230560

13. Zhao, X. Leveraging Artificial Intelligence (AI) technology for English writing: Introducing Wordtune as a digital writing assistant for EFL writers. *RELC J*. 2022;54(3):890-894.

    doi: 10.1177/00336882221094089

14. Abdullahi A. *10 Best AI Writing Tools 2023*. eWEEK; 2023. Available from: https://www.eweek.com/artificial-intelligence/ai-writing-tools [Last accessed on 2024 Jul 09].

15. Pividori M, Greene CS. A publishing infrastructure for AI-assisted academic authoring. *bioRxivorg*. 2023.

    doi: 10.1101/2023.01.21.525030

16. Cabanac G, Labbé C. Prevalence of nonsensical algorithmically generated papers in the scientific literature. *J Assoc Inf Sci Technol*. 2021;72(12):1461-1476.

    doi: 10.1002/asi.24495

17. Scimeca M, Bonfiglio R. Dignity of science and the use of ChatGPT as a co-author. *ESMO Open*. 2023;8(4):101607.

    doi: 10.1016/j.esmoop.2023.101607

18. Mullin B, Grant N. *Google Tests A.I. Tool that is Able to Write News Articles*. The New York Times; 2023. Available from: https://www.nytimes.com/2023/07/19/business/google-artificial-intelligence-news-articles.html [Last accessed on 2024 Jul 09].

19. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber Phys Syst*. 2023;3:121-154.

    doi: 10.1016/j.iotcps.2023.04.003

20. Megawati R, Listiani H, Pranoto NW, Akobiarek M, Megahati SRR. Role of GPT chat in writing scientific articles: A systematic literature review. *JPPIPA*. 2023;9(11):1078-1084.

    doi: 10.29303/jppipa.v9i11.5559

21. Stokel-Walker C. ChatGPT listed as author on research papers: Many scientists disapprove. *Nature*. 2023;613(7945):620-621.

    doi: 10.1038/d41586-023-00107-z

22. Hosseini M, Rasmussen LM, Resnik DB. Using AI to write scholarly publications. *Account Res*. 2023:1-9.

    doi: 10.1080/08989621.2023.2168535

23. Park JY. Could ChatGPT help you to write your next scientific paper?: Concerns on research ethics related to usage of artificial intelligence tools. *J Korean Assoc Oral Maxillofac Surg*. 2023;49(3):105-106.

    doi: 10.5125/jkaoms.2023.49.3.105

24. Else H. Abstracts written by ChatGPT fool scientists. *Nature*. 2023;613(7944):423.

    doi: 10.1038/d41586-023-00056-7

25. Thorp HH. ChatGPT is fun, but not an author. *Science*. 2023;379(6630):313.

    doi: 10.1126/science.adg7879

26. Dergaa I, Chamari K, Zmijewski P, Ben Saad H. From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biol Sport*. 2023;40(2):615-622.

    doi: 10.5114/biolsport.2023.125623

27. Sharma P. Chatbots in medical research: Advantages and limitations of artificial intelligence-enabled writing with a focus on ChatGPT as an author. *Clin Nucl Med*. 2023;48(9):838-839.

    doi: 10.1097/RLU.0000000000004665

28. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*. 2023;15(2):e35179.

    doi: 10.7759/cureus.35179

29. Dashti M, Londono J, Ghasemi S, Moghaddasi N. How much can we rely on artificial intelligence chatbots such as the ChatGPT software program to assist with scientific writing? *J Prosthet Dent*. 2023.

    doi: 10.1016/j.prosdent.2023.05.023

30. Masic I, Miokovic M, Muhamedagic B. Evidence based medicine - New approaches and challenges. *Acta Inform Med*. 2008;16(4):219-225.

    doi: 10.5455/aim.2008.16.219-225

31. Ansari M. *Impact of Online Reading on Skills of Professionals. Library Philosophy and Practice* (*E-Journal*). Pakistan: University of Karachi; 2018. Available from: https://digitalcommons.unl.edu/libphilprac/1753.html [Last accessed on 2024 Jul 09].

32. ChatGPT. *Openai.com*. Available from: https://chat.openai.com [Last accessed on 2024 July 16].

33. OpenAI. *GPT-4 Technical Report*. arXiv [csCL]; 2023. Available from: https://arxiv.org/abs/2303.08774 [Last accessed on 2024 Jul 16].

34. Rahaman MS, Ahsan MMT, Anjum N, Terano HJR, Rahman MM. From ChatGPT-3 to GPT-4: A significant advancement in AI-driven NLP tools. *J Eng Emerg Technol*. 2023;2(1):50-60.

    doi: 10.52631/jeet.v2i1.188

35. Koubaa A. GPT-4 vs. GPT-3.5: A concise showdown. *Preprints*. 2023.

    doi: 10.20944/preprints202303.0422.v1

36. Ahmad FB, Cisewski JA, Xu J, Anderson RN. Provisional mortality data - United States, 2022. *MMWR Morb Mortal Wkly Rep*. 2023;72(18):488-492.

    doi: 10.15585/mmwr.mm7218a3

37. *World Health Statistics 2020: Monitoring Health for the SDGs, Sustainable Development Goals*. Geneva: World Health Statistics. Available from: https://www.who.int/publications/i/item/9789240005105 [Last accessed on 2024 Jul 16].

38. *Health Statistics at Regional Level*. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=health_statistics_at_regional_level [Last accessed on 2024 Jul 09].

39. Home. *SCI Journal*. Available from: https://www.scijournal.org [Last accessed on 2024 July 16].

40. Li J, Cheng X, Zhao X, Nie JY, Wen JR. *HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models*. arXiv [csCL]; 2023. Available from: https://arxiv.org/abs/2305.11747 [Last accessed on 2024 Jul 16].

41. Zhang Y, Li Y, Cui L, Cai D, Liu L. *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. arXiv [csCL]; 2023. Available from: https://arxiv.org/abs/2309.01219 [Last accessed on 2024 Jul 16].

42. US Census Bureau. *Frequently Occurring Surnames from the 2010 Census*; 2019. Available from: https://www.census.gov/topics/population/genealogy/data/2010_surnames.html [Last accessed on 2023 Dec 13].

43. Jiang Z. *公安部发布去年全国姓名报告，"王、李、张"姓排前三* [*Public Security Bureau Releases Last Year's National Name Report, "Wang, Li, Zhang" Ranked Top Three*]. *澎湃新闻* [*The Paper*]; 2019. Available from: https://www.thepaper.cn/newsDetail_forward_2927573 [Last accessed on 2023 Dec 13]. [In Chinese]

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Innovative infrared imaging approach for breast cancer screening: Integrating rotational thermography and machine learning analysis

Asok Bandyopadhyay[1†*], Himanka S. Mondal[1†], Bivas Dam[2], Dipak C. Patranabis[2], and Barnali Pal[1]

[1]ICT&SERVICES Group, Centre for Development of Advanced Computing, Kolkata, West Bengal, India

[2]Department of Instrumentation and Electronics, Jadavpur University, Kolkata, West Bengal, India

Abstract

This paper presents a novel approach to breast cancer screening using infrared (IR) imaging. This work encompasses four phases: Refining data collection, advancing analysis methods, and enhancing feature extraction with machine learning. The developed system employed a temperature-controlled chamber with rotational thermography techniques to maintain consistent temperatures and capture high-quality IR images and all possible subject views. The paper describes four key experiments to detect breast cancer using IR imaging. The experiments involved the use of dynamic temperature-based data collection and a semi-circular arc movement to ensure precise imaging, keeping the object in focus. Initial experiments involved the use of dynamic temperature-based data collection and a semi-circular arc movement to ensure precise imaging focus. The final experiment incorporated a semi-circular arc movement. For each subject, 32 thermal IR images were acquired, targeting one breast at a time while isolating the other with an IR-proof barrier. The collected datasets were used for breast abnormality detection. The analyzed results revealed that support vector machine and neural network algorithms achieved an accuracy rate of 93.18%. The system's installation at a hospital in India allowed for real-world application and validation. The final study, which introduced a new IR imaging protocol, demonstrated improved results compared to earlier pilot studies. This method enhances the accuracy of distinguishing malignant and benign tumors, supporting early breast cancer detection and treatment. The proposed methodology addresses data collection and analysis challenges, leading to improved screening efficiency and better patient outcomes.

Keywords: Infrared technology; Thermal imaging; Breast cancer screening; Dynamic data; Data collection methods; Rotational thermography

†*These authors contributed equally to this work.*

***Corresponding author:**
Asok Bandyopadhyay
(b_ashoke@hotmail.com)

Citation: Bandyopadhyay A, Mondal HS, Dam B, Patranabis DC, Pal B. Innovative infrared imaging approach for breast cancer screening: Integrating rotational thermography and machine learning analysis. *Artif Intell Health*. 2024;1(3):64-79.
doi: 10.36922/aih.3312

Received: March 28, 2024

Accepted: May 24, 2024

Published Online: July 23, 2024

1. Introduction

Infrared (IR) technology-based imaging has emerged as a versatile medical imaging technique, facilitating the capture of temperature distributions across the human body's surface. IR imaging has gained attention in medical diagnosis by complementing other imaging methods. Notably, studies have highlighted a correlation between malignant

breast tumors and elevated temperatures, prompting consideration of IR imaging as a valuable tool for breast cancer detection.

In recent years, the global rise in breast cancer incidence has spurred extensive research into IR imaging for breast cancer diagnosis. Prior research has explored thermal imaging for various medical applications, including diabetic foot disease, ocular surface temperature analysis, and tumor detection[1-6] Click or tap here to enter text. Click or tap here to enter text. Click or tap here to enter text. Click or tap here to enter text. Click or tap here to enter text. Studies investigating breast cancer detection through thermal imaging have examined bi-spectral invariant features, color segmentation techniques, and deep learning-based segmentation methods[7-14] Click or tap here to enter text. Click or tap here to enter text. Click or tap here to enter text. Click or tap here to enter text. Click or tap here to enter text. Click or tap here to enter text. Click or tap here to enter text. A novel method combining rotational thermography and dynamic temperature analysis has been developed to address challenges in breast cancer screening[15-17] Click or tap here to enter text. Click or tap here to enter text. This non-contact, non-invasive approach enhances the visibility of abnormalities in breast tissue, improving detection accuracy. Machine learning algorithms trained on extracted features aid in distinguishing normal from abnormal patterns. Earlier, EtehadTavakol *et al.*[7,10] achieved high detection rates using bi-spectral invariant features and K-means clustering for segmentation.[7,10] Garduño-Ramón *et al.*[11] introduced a non-invasive tool utilizing temperature and texture features, yielding promising results[11] Various segmentation techniques, including K-means, fuzzy c-means (FCM), and expectation-maximization (EM) algorithms, have been explored, with the EM algorithm demonstrating superior accuracy[11,12] Click or tap here to enter text. Venkataramani *et al.*[10,13,18-22] proposed a semi-automated method using morphological filtering and thresholding, achieving high sensitivity and specificity.[10,13,18-22] Deep learning-based approaches, such as the level-set method, have shown high accuracy in segmenting suspicious regions in breast thermograms.[9,23,24] Thermal imaging has also been utilized for brain tumor detection, highlighting its versatility in medical diagnostics.[13,20,25-27]

While IR imaging shows promise for breast cancer screening, challenges remain, including standardization of temperature values and the need for trained personnel. IR Imaging researches also focus on optimizing machine learning algorithms, integrating advanced imaging techniques, and exploring novel optimization algorithms for enhanced diagnosis and classification.[28-34]

However, one of the primary challenges lies in establishing precise temperature thresholds to differentiate malignant tumors amidst variations in individual heat sources due to diverse medical and physical conditions. Addressing this challenge is crucial in developing effective breast cancer screening systems. This study presents a novel, non-contact, non-invasive breast imaging method capable of capturing comprehensive abnormalities. Several obstacles confront researchers in the development of IR imaging-based breast cancer screening systems, including the surge in breast cancer cases, the cost and invasiveness of current screening modalities, challenges in tracking malignancy progression, poor visibility of affected anatomical areas, patient discomfort during examinations, and shortages of skilled medical professionals. Table 1 summarizes the earlier works and comparison with the approach used in this study. This research endeavors to overcome these challenges through an efficient data collection process and systematic stages of system development for IR image processing techniques. Conducted at a prominent hospital in northeast India, the study comprises four research phases: Three pilot studies (Phase 1 [PS1], Phase 2 [PS2], and Phase 3 [PS3]) and a final study (FS). The study evolves to optimize imaging quality by transitioning from low-resolution forward-looking IR (FLIR) SC-325 cameras in the initial phases to Infratec HD 600 and FLIR T-650 cameras in subsequent phases. The study aimed to develop an efficient and accurate thermal imaging system for breast cancer screening. It progressed through four phases, refining data collection methods and improving imaging results.

In PS1 and PS2, conventional imaging approaches were employed, but challenges such as varied focal lengths and patient movement resulted in unsatisfactory images. PS3 introduced a semi-circular arc for camera movement, allowing for precise focus on one breast at a time, though manual adjustments and open environments presented difficulties. Findings from PS3 and FS underscore the necessity of mounting the IR camera on a motorized mechanical arm capable of rotational thermography, enabling comprehensive imaging coverage. In the final phase (FS), a temperature-controlled chamber and automated hardware and software provided a standardized, touchless, and painless imaging process with an accuracy rate of 93.18% for detecting abnormalities. Data analysis evolved from conventional methods to advanced techniques in PS3 and FS, employing clustering methods and references from ultrasonography (USG) and biopsy reports. Mean temperature and standard deviation analyses further enhance detection precision. The study included diverse breast cancer subtypes and stages, demonstrating the system's applicability across various scenarios. Control groups provided benchmarks for assessing accuracy and

Table 1. A summary table of earlier works and a comparison with our approach

Study	Findings	Pros	Cons	Uniqueness of our study
Etehad Tavakol *et al.*[7,10]	High detection rates using bi-spectral invariant features and K-means clustering for segmentation.[7,10]	High detection rates; effective segmentation technique.	Limited scope; may lack versatility across different conditions.	Our study expands on bi-spectral invariant features with more advanced machine learning algorithms.
Garduño-Ramón *et al.*[11]	Non-invasive tool utilizing temperature and texture features, yielding promising results.[11]	Non-invasive; promising outcomes.	May require further optimization for diverse cases.	Our study includes more precise temperature controls and advanced segmentation techniques.
Various segmentation techniques	K-means, fuzzy c-means, and EM algorithms were explored, with EM showing superior accuracy.[11,12]	Superior accuracy with EM algorithm.	May face challenges in real-time applications.	Our study integrates advanced techniques, including fuzzy c-means clustering, for better real-time outcomes.
Venkataramani *et al.*[10,13,18-22]	Semi-automated method using morphological filtering and thresholding, achieving high sensitivity and specificity.[10,13,18-22]	High sensitivity and specificity.	Semi-automated methods may still require human intervention.	Our study uses a fully automated approach with robotic arm movement and data processing.
Deep learning-based approaches	High accuracy in segmenting suspicious regions in breast thermograms.[9,23,24]	High accuracy in segmentation.	Deep learning models may require large datasets for training.	Our study applies machine learning in tandem with a novel data collection protocol for more comprehensive results.
General challenges in IR imaging	Standardization of temperature values; need for trained personnel; variability in individual heat sources.[28-34]	Promising results in detecting abnormalities.	Difficulty in establishing precise temperature thresholds.	Our study introduces a non-contact, non-invasive approach with precise temperature control for improved results.

Abbreviations: EM: Expectation-maximization; IR: Infrared.

specificity. Neural network (NN) parameters and pattern recognition tools assessed the system's performance with high accuracy rates across different phases.

2. Data and methods

2.1. Data collection

In PS1, data collection was undertaken following the technique previously described in the literature.[14,35] The subject was seated in front of the camera with both hands raised upward, as illustrated in Figure 1A. In PS2, the subject was seated similarly to PS1 with two cameras deployed at two corners for IR image acquisition.[14,30,36] Figure 1B illustrates the setup for the same.

Several modifications were carried out in the imaging process based on the doctors' guidance. One modification was placing a camera in a fixed position with the patient seated on a rotating chair. In this case, images were acquired from various predetermined angles. The data collection angles are 0°, 30°, 60°, 90°, 120°, 150°, and 180° from the initial position. Since the patient was rotating, focusing on a specific breast was challenging. It also led to a shifting region of interest (ROI) in the images. This variation in focal length led to unsatisfactory imaging results.

The subsequent logical adaptation was repositioning the camera while keeping the patient stationary. However, a significant challenge arose in obtaining a clear IR image of the breast from a particular angle, as the camera movement and rotation caused the images to overlap. This resulted in the inner quadrant of one breast's IR image being superimposed by the other breast's outer quadrant.

Accordingly, a significant modification was made to address this concern in PS3, as illustrated in Figure 1C. The patient was seated in a fixed position, and the camera moved in a semi-circular arc on an arm-based arrangement. A tabletop mechanical arrangement was developed to ensure that only one breast was focused at the pivot point of the semi-circular arc. The second breast was isolated from the camera view by covering it with an IR-proof barrier. A tabletop setup of PS3 is shown in Figure 2.

Finally, rotational thermography was set up in the FS. A camera rotated in a semi-circular arc and stopped at different angles, with the patient seated in a fixed position. The arrangement is illustrated in Figure 3.

It comprises an enclosed chamber with a breast-shaped grooved hole through the chamber wall. The patients' breasts are positioned one at a time through this hole for IR imaging. The distance from the camera to the subject is 1 m, the minimum focus distance of the IR camera calibrated by the manufacturer at a thermal laboratory. The ground clearance of the system was 0.76 m. A total of 32 thermal IR images were acquired for each subject, with 16 images acquired at a higher ambient temperature, for example,

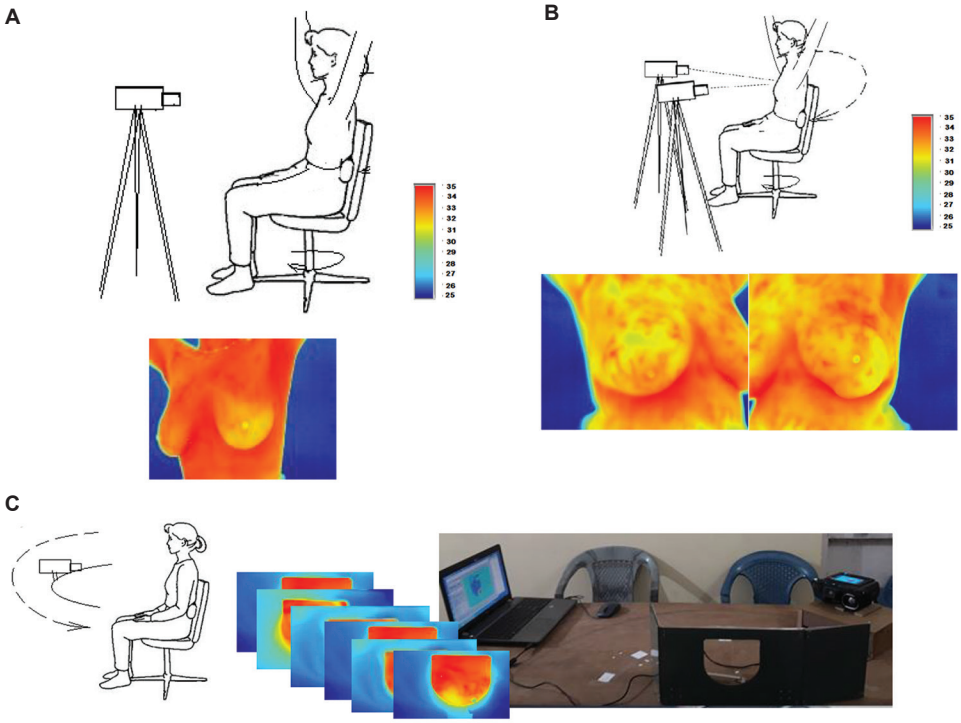


Figure 1. Illustration of different patient sitting positions during phases. (A) Data acquisition in Phase 1 with a sample infrared (IR) image. Infrared images were collected using a single IR camera while the patient sat on a rotating chair. (B) Data acquisition in Phase 2 with a sample IR image collected using a dual infrared camera while the patient sat on a rotating chair. (C) Data acquisition setup in Phase 3. Infrared images were collected using a single IR camera while the patient sat on a chair. The camera setup rotated on a tabletop setup. Note: Infrared images shown here were collected during data collection. Illustrations were created using MS Paint.

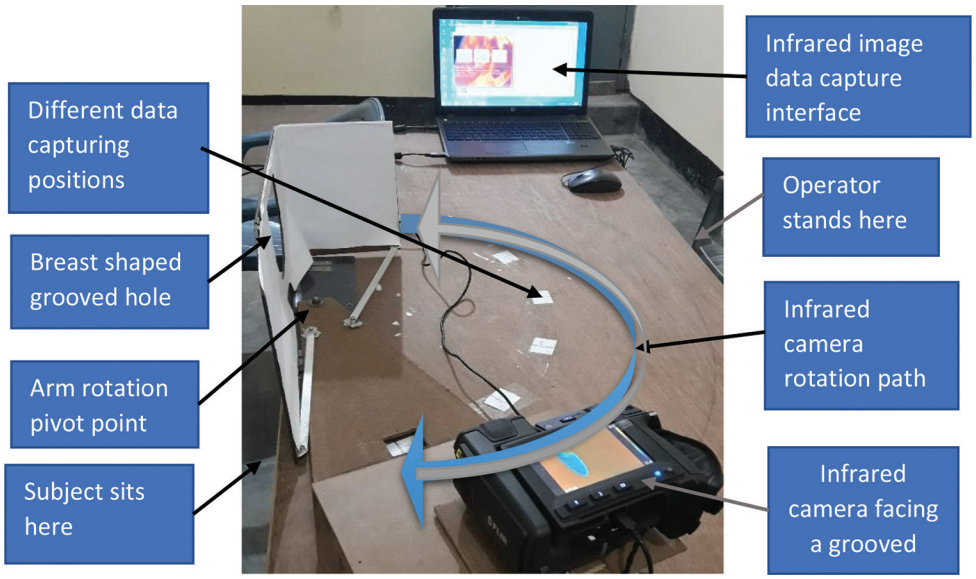


Figure 2. Table-top setup for rotational thermography (Phase 3). The setup images shown here were collected during data collection.

25°C, and the other 16 at a lower ambient temperature, for example, 23°C. The temperature variation between the two ambient conditions was 2°C. The IR images were acquired every 30° from 0 to 180°.

Initially, 14 IR images (seven for each breast) and two from the axilla (one from each side) were taken from each IR image set at a particular ambient temperature. During data collection, 17 steps were taken to position and focus

the camera at a particular angle to capture the required IR breast images. These steps are detailed in Table 2. After allowing the ambient temperature to reduce by 2°C through the temperature controller of the air conditioner (AC), the second set of data was acquired. Accordingly, 32 IR breast images were collected over both ambient temperatures. This method is termed dynamic temperature-based data collection.

Protocol-wise, the room temperature is adjusted for each subject in the FS. Controlling the room's temperature and maintaining it at a specific value effectively and quickly was difficult for each patient. The tabletop rotating mechanical setup placed the IR image-capturing system in an airtight chamber to overcome this problem. The temperature inside this chamber was externally regulated through a portable AC. The room temperature was maintained at a constant value through the room AC. This setup is shown in Figure 4.

2.2. Data analysis

In thermal imaging, the camera is the primary device for acquiring the data to be analyzed for breast cancer screening. While previous studies are based on this methodology, their technique is different and not in consonance with the imaging system to generate appropriate images considering variations in temperature and other physical conditions that influence imaging. Hence, detailed knowledge and appropriate system adaptation are essential. Without such intelligent data collection processes, data analysis is crucial for algorithmic perfection and improved machine learning performance.[9,11] Therefore, integrating data collection and analysis while exploring the impact of environmental conditions on IR imaging is essential and required at every step for optimal system performance.

While some researchers have developed data collection protocols, the data analysis algorithms are unsuitable for integration with IR image processing techniques, resulting in poor execution for the breast cancer screening system. This study integrated IR image processing techniques into the system at every stage, leading to efficient system development and gradual software evolution. This results in an effective IR image processing algorithm. Previous studies used online databases like Thermalytix[37] to develop IR imaging algorithms. However, its real-life implementation had a limitation as this database needs more information for such development. Therefore, a novel data collection protocol was developed and implemented in the FS to overcome this limitation.

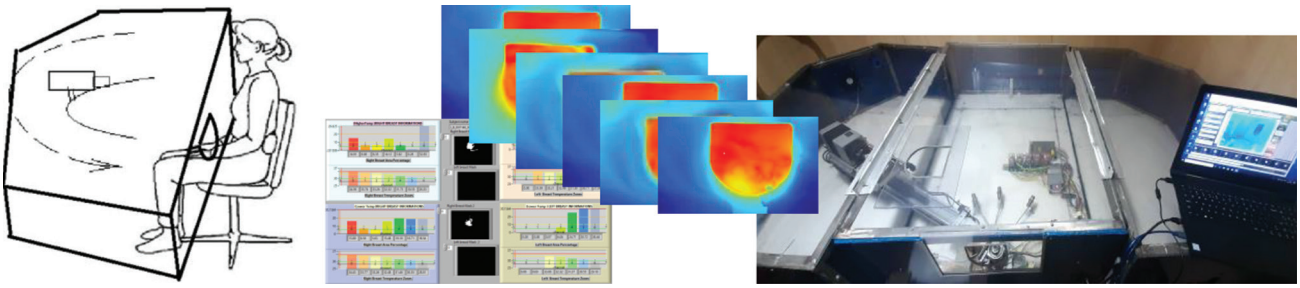


Figure 3. Rotational thermography setup in the final study consisting of an infrared camera in a temperature-controlled enclosure. Infrared images were collected using a single infrared camera while the patient sat on a chair. The camera setup rotated on a tabletop setup that was enclosed by a temperature-controlled chamber. The infrared images shown here were collected during data collection. Illustrations were created using MS Paint

Table 2. Details of the research phases and evolution to the final stage

Phase	Description	No. of patients	Design setup	Challenges/observations
PS1	Two low-resolution cameras were used for image acquisition from 45° angles.	71	Subject is seated in front of the cameras with raised hands.	Skilled personnel are required for data collection and analysis.
PS2	A single camera was fixed, and images were acquired from various angles while the patient was seated on a rotating chair.	10	Fixed camera setup with a rotating chair.	Challenging to focus on a specific breast due to patient rotation.
PS3	Exploration of different positioning techniques in an open space air-conditioned room.	33	Semi-circular camera movement with one breast focused and the other covered.	Manual hardware movement and difficulty in controlling ambient temperature.
FS	Rotational thermography in a temperature-controlled chamber.	88	Rotating camera in a semi-circular arc.	Automated hardware movement and standardized protocol in an enclosed environment.

Abbreviations: FS: Final study; PS: Phase.

2.3. IR-image feature-based analysis technique

In PS1 and PS2, the IR images were analyzed conventionally, following the method adopted in previous studies. The features extracted were the mean, median, mode, standard deviation, histogram, and maximum value. Analysis was conducted in consultation with doctors, but no reference was made to IR images acquired through USG, mammography, or biopsy.

In PS3, the primary reference source was the USG and biopsy reports obtained through other modalities. IR image-based clustering was used for image segmentation and to extract the ROI.

The mean temperature of each ROI, interpreted as different body temperature zones, was used as the discriminating feature. IR image K-means clustering was used for clustering the other image features.[12,25,28] The clustering method was gradually improved, and the number of clusters was optimized from 20 to seven based on experimental validation by consulting doctors based on abnormalities found in the USG and biopsy reports. This study extracted temperature-based clustering features for IR image segmentation for 33 subjects.[8] In the next stage of development (FS), the image background and foreground were separated through FCM clustering.[7] Figure 5 shows the variation in the IR breast images captured from different angles, with the abnormality detected by the software and doctors as irregular and box-shaped ROI, respectively.

Final IR image analysis was accomplished after integrating the temperature-controlled enclosure into the system. The higher ambient temperature state was chosen as the reference temperature. The ROI was divided into seven clusters through K-median clustering and was defined as the features in the machine learning algorithm for processing. Seven clusters were finalized based on experimental validation of the abnormality found in the USG and biopsy reports by the consulting doctors. Subsequently, another dataset was acquired in the lower ambient temperature state. The differences between the two datasets were the key discriminating features for analysis. This novel analysis technique was tested on 88 subjects from a hospital in northeast India. Figure 6 displays the frontal view of the IR images of a subject's affected and normal breasts, highlighted by the red and blue zones, respectively.

2.3.1. Temperature area clustering method

The number of pixels corresponding to each temperature cluster zone was recorded during IR image analysis. The total number of pixels represented the area of each zone. For example, the camera used in this study captures a 640 × 480–pixel IR image, which is considered to be 100% area. Accordingly, if a particular zone had 7962 pixels, it spread over 2.59% of the IR image and was known as a percent area cluster. The distributions of these area zones across different temperatures are depicted in Figure 7 using bar plots. Higher temperature regions indicative of abnormalities are conventionally represented in specific colors. The color scale in the figure does not mark the seven cluster zones described in this paper. Instead, it only demonstrates the color temperature relative to the IR image, while the clusters are outlined plots overlayed on the IR image.

Figure 7 illustrates the real-world implications of the temperature zone versus percent area cluster analysis. For the patient in question, imaging at a higher ambient temperature revealed that the right breast had 14.91% of its area at 34.59°C (zone 0). When the ambient temperature was lowered, the area of the right breast at 34.43°C increased to 15.89% (zone 0). Since the highest body temperature did not decrease with a change in ambient temperature, we concluded that the right breast had an abnormality. Conversely, for the left breast, the highest temperature at a higher ambient temperature was 33.50°C, covering 0.59%

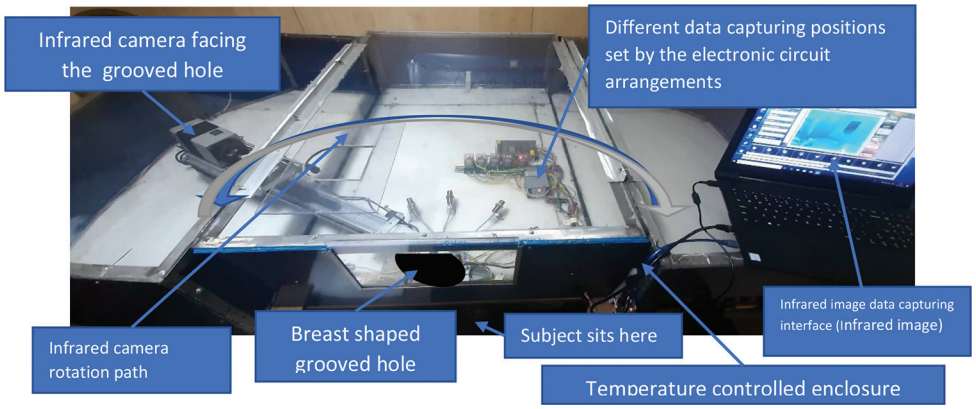


Figure 4. Rotational thermography setup in a temperature-controlled enclosure. Setup images shown here are collected during data collection

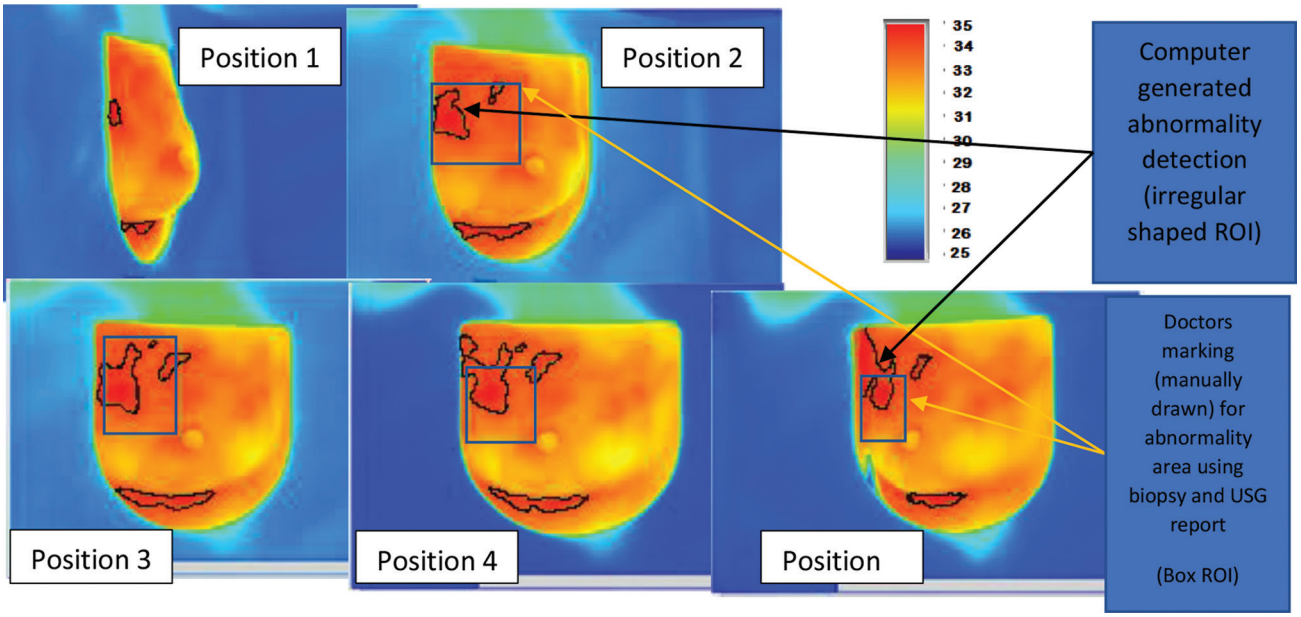


Figure 5. Region of interest of detected abnormality on infrared breast images captured from different angles. Infrared images shown here were collected during data collection. Illustrations were created using MS Paint.
Abbreviation: USG: Ultrasonography.

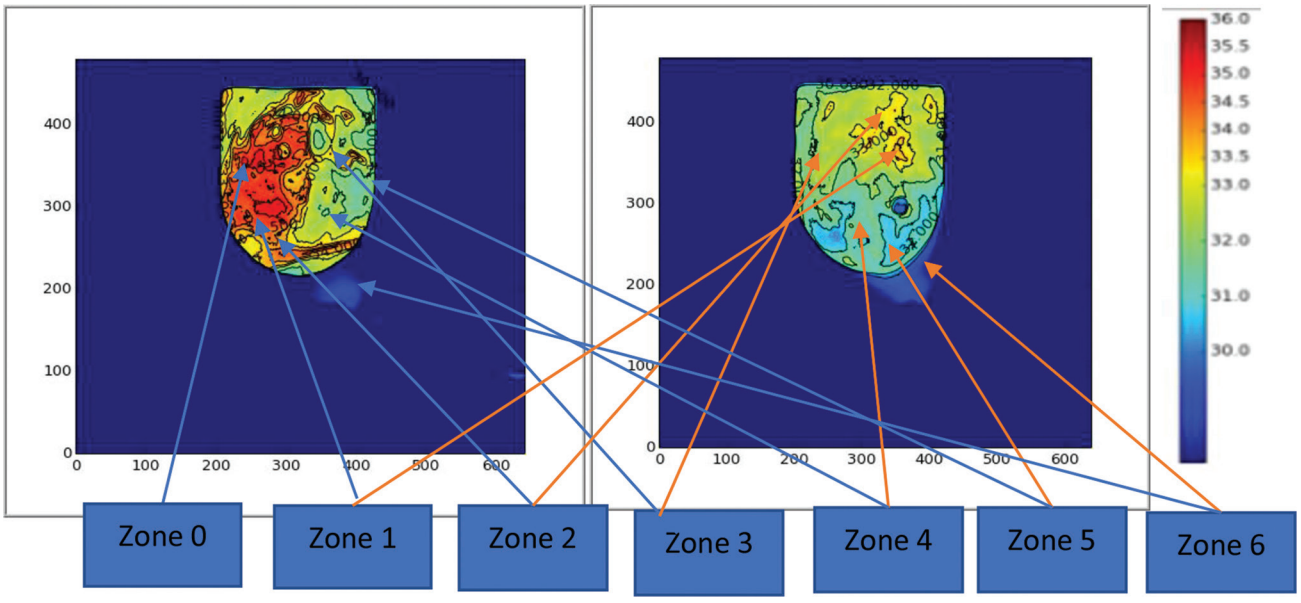


Figure 6. Frontal view of the infrared images of the affected (right) and normal (left) breast. Infrared images shown here were collected during data collection. Illustrations were created using MS Paint.

of the area (zone 1). While the ambient temperature was lowered, 0% of the area was at that temperature (zone 1). Instead, the highest temperature was shifted to zone 2, which dropped to 0.70% of the area. This implies that the body temperature of the left breast significantly decreased with a change in ambient temperature, implying no abnormality.

The temperature area clustering method and its related algorithm were implemented through LabVIEW, which has been copyrighted. Utilizing the LabVIEW environment for image analysis and validation through clinical summary reports, this study introduces an intelligent data collection protocol to enhance system performance, building upon previous research. Ethical approval from the Cachar Cancer

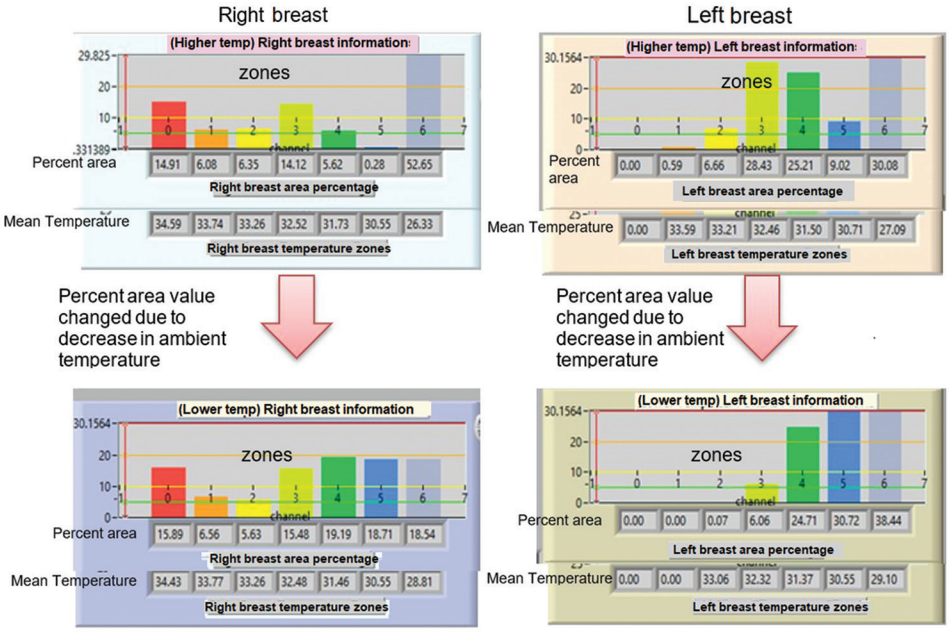


Figure 7. Percent area and temperature zone for two ambient temperatures. Illustrations were created using MS Paint.
Abbreviation: Temp: Temperature.

Hospital and Research Centre IRB/Ethics Committee ensures adherence to rigorous ethical standards in research conduct.

3. Results and discussion

3.1. Result A

The study aimed to develop an efficient and accurate imaging system for breast cancer screening using thermal imaging techniques. Four phases were conducted to refine the data collection process and improve imaging results. The challenges and observations encountered in each phase led to the development of a more advanced and reliable system in the FS.

In PS1, data acquisition utilized two low-resolution FLIR cameras. The patients were seated in front of the cameras with both hands raised upward. Fixed angulated images were taken from a 45° angle. Skilled personnel were required for data collection and analysis, making it a human-dependent process. This approach provided initial insights into the use of thermal imaging for breast cancer screening.

In PS2, a similar setup to PS1 was used, but with one FL SC325 camera, and angular views were captured. The camera was fixed while the patient sat on a rotating chair, and images were acquired from various predetermined angles. However, the patient's rotational movement made it challenging to focus on a specific breast, leading to variations in focal length and unsatisfactory imaging results.

To address these challenges, PS3 introduced rotational thermography. The patient was seated in a fixed position, and the camera moved in a semi-circular arc on an arm-based arrangement. A tabletop mechanical setup ensured that only one breast was focused at the arc's pivot point while the other breast was covered with an IR-proof barrier. This modification improved control and focus on a single breast, improving imaging results. However, manual hardware movement and data collection in an open-space environment pose difficulties in maintaining reliable and standardized data.

Finally, in the FS phase, a temperature-controlled chamber was introduced to overcome the challenges faced in previous phases. The camera was set to rotate in a semi-circular arc and stop at various angles while the patient remained fixed. The enclosed chamber provided a controlled environment for data collection, eliminating the need for manual temperature control. Thirty-two thermal IR images were captured for each subject at higher and lower ambient temperatures. This phase employed robotic arm-based automated hardware movement and an automated software analysis tool, resulting in a standardized protocol and improved data collection system.

The proposed imaging system showcased several salient features, including a touchless and painless process for maximum patient comfort, a rotating gantry for acquiring

different angles of the breast, dynamic IR image collection in a temperature-controlled enclosed chamber, and a simplified user interface for data collection, analysis, and interpretation.

Regarding data analysis techniques, PS1 and PS2 followed conventional methods used in previous studies. Features such as mean, median, mode, standard deviation, histogram, and maximum value were extracted from the IR images. However, no reference was made to images acquired through other modalities, such as USG, mammography, or biopsy. The PS3 and FS are complete with the information about PS1 and PS2, assessed for their value in the research article.

In PS3, a shift toward a more comprehensive analysis technique was observed. The primary reference source became the USG, and biopsy reports were obtained through other modalities. IR image-based clustering was employed to segment and extract the ROI. The mean temperature of each ROI was used as a discriminating feature, and K-means clustering was applied to cluster other image features. The clustering method was gradually improved, and the number of clusters was optimized based on experimental validation and consultation with doctors.

The FS phase enhanced the data analysis technique by separating the image background and foreground through FCM clustering. This allowed for better detection and delineation of abnormalities in the IR breast images. The collaboration between software analysis tools and medical experts resulted in identifying irregularly shaped and box-shaped ROIs as potential abnormalities, as shown in Figure 5.

The study demonstrated the importance of integrating data collection and analysis techniques to improve the performance of breast cancer screening systems using IR thermography. The iterative improvements in the data acquisition setup and analysis algorithms led to a more standardized and efficient process. With its temperature-controlled chamber, rotating camera setup, and automated analysis tools, the FS phase showed promising results for accurate and reliable breast cancer screening.

The data presents a comprehensive overview of breast cancer subtypes, sizes, and stages within the examined cohort, providing valuable clinical context. The subtype's distribution includes 63% for ductal carcinoma *in situ*, 25% for invasive ductal carcinoma, 8% for invasive lobular carcinoma, and 4% for other subtypes. Cancer sizes ranged from 0.8 cm to 5.6 cm, with an average size of 2.3 cm. Regarding cancer stages, 42% were Stage I, 33% were Stage II, 15% were Stage III, and 10% were Stage IV. These insights into the cohort's diversity enhance the system's applicability across various breast cancer scenarios. Table 3 displays the patient data table with all the patients' demographics and the disease stage.

In terms of control, the study employed a comprehensive diagnostic approach, including comparisons between symptomatic and asymptomatic patients, various imaging techniques, and cross-referencing with other diagnostic methods. This control approach thoroughly assessed the system's performance and reliability in different clinical contexts.

Healthy patients were recruited as controls, providing a benchmark for assessing the system's accuracy and specificity. Their demographics included a diverse range of ages and breast health statuses, reflecting the general population and allowing for a comprehensive evaluation of the system's performance across different patient profiles.

The correlation between temperature and readouts is critical for evaluating the breast cancer screening system using IR thermography. Analyses of mean temperature and standard deviation showed the system's ability to identify potential abnormalities based on temperature variations in breast tissue. Improved clustering techniques in later phases enhanced the precision of detecting abnormalities. Overall, the positive correlation between temperature and readouts highlights the system's potential as an effective diagnostic tool for breast cancer screening.

3.2. Result B

A meticulous qualitative statistical analysis evaluated abnormality detection accuracy in the captured IR images. Table 4 comprehensively summarizes the detected abnormalities and their corresponding accuracy percentages for each study phase. Notably, a progressive enhancement in accuracy was observed throughout the study duration, with PS3 and the final stage (FS) achieving

Table 3. Patient data table with all the demographics of the patients and the stage of disease

Category	Percentage	Details
Subtypes	63	Ductal carcinoma *in situ*
	25	Invasive ductal carcinoma
	8	Invasive lobular carcinoma
	4	Other subtypes
Tumor sizes	Range	0.8 – 5.6 cm
	Average	2.3 cm
Stages	42	Stage I
	33	Stage II
	15	Stage III
	10	Stage IV

Table 4. Comparison of studies in the population-based case–control study

Phase	Data collection	Number of subjects	Mode	TP	TN	FP	FN	Sensitivity (%)	Specificity (%)	Accuracy (%)
PS1	During June 2015 – June 2016	71	Manual	This study was not done on preliminary collected dataset						
PS2	During June 2016 – Aug 2016	10	Manual	1	6	2	1	50%	75%	70%
PS3	October 2017	33	Semi-automated	9	19	4	1	90.00	82.61	84.84
FS	November 2017 – September 2019	88	Automated	23	59	1	5	82.14	98.33	93.18

Abbreviations: EM: Expectation-maximization; FN: False negative; FP: False positive; FS: Final study; IR: Infrared; PS: Phase; TN: True negative.

notably higher detection rates compared to the initial phases (PS1 and PS2). These findings underscore the efficacy of the developed system in identifying potential abnormalities indicative of breast cancer.

In PS1, involving a cohort of 71 patients, the system detected 23 abnormalities. This phase used a manual evaluation process, and a case–control-based study was not conducted. In PS2, with a smaller sample size of 10 patients, the system detected five abnormalities, resulting in an accuracy rate of 70%. The sensitivity in this phase was 50%, meaning half of the actual positives were correctly identified, and the specificity was 75%, indicating a better performance in correctly identifying true negatives. The calculated area under the receiver operating characteristic curve (AUC) for this phase is 0.625. Despite the reduced sample size, the system demonstrated promise in identifying abnormalities.

PS3 marked a significant advancement, encompassing 33 patients. The system detected 30 abnormalities, yielding an impressive accuracy rate of 84.84%. The sensitivity was 90%, highlighting the system's enhanced ability to identify actual positives correctly. The specificity improved to 82.61%, indicating fewer false positives. These metrics underscore the advancements in positioning techniques and enhanced image quality. The calculated AUC for this phase is 0.863, showcasing significant overall discriminative power in distinguishing between positive and negative cases. Phase 4 (FS), involving 88 patients, exhibited the highest performance with an accuracy rate of 93.1%. The sensitivity was 82.14%, showing a high true positive rate, and the specificity reached 98.33%, indicating an excellent true negative rate. These results underscore the system's capability to identify potential abnormalities indicative of breast cancer. This phase illustrates the system's reliability and effectiveness, providing a promising tool for early detection and enhancing patient outcomes. The calculated AUC for this phase is 0.902, further highlighting the system's discriminative power in distinguishing between positive and negative cases.

The study ensured robust model evaluation by dividing the data into distinct phases (PS3 and FS) for training, validation, and testing. In PS3, the training set consisted of 23 subjects (368 images), with five subjects (80 images) each for validation and testing. For FS, 62 subjects (992 images) were used for training, while 13 subjects (208 images) each were allocated to validation and testing. This approach allowed for efficient model tuning and performance evaluation on unseen data, improving the model's reliability. The results of the NN tools generated are shown in Figures 8 and 9.

The NN parameters included the distribution of different area zones corresponding to different temperatures. Specifically, the study recorded the number of pixels corresponding to each temperature cluster zone, using this as the area of that zone. The IR images (640 × 480 pixels) were considered 100% area, and the NN analyzed how different area zones were distributed across the images. The network's parameters included seven zones in two ambient temperatures across 16 images per subject. The NN parameters typically included the number of Layers: Three hidden layers; learning rate: 0.001; optimization function: Adam optimizer. Pattern recognition tools were used for image classification and assessed through confusion matrices for accuracy and performance validation. The study employed 5-fold cross-validation as the validation technique for machine learning algorithms, ensuring a thorough and reliable evaluation of the model's performance. This method involves dividing the data into five subsets and training the model 5 times, each using a different subset as the validation set and the remaining subsets as the training set. This approach allowed for a comprehensive evaluation of the NN's ability to identify abnormalities.

Furthermore, mean temperature and standard deviation analysis were conducted to assess the stability and variation of temperature measurements across different phases. Mean temperature was calculated as the sum of individual temperatures divided by the total number of subjects, providing insights into temperature stability. On the other hand, standard deviation offered valuable information regarding temperature variation within the dataset.

The confusion matrices in Figures 8 and 9 show the classification outcomes of the NN pattern recognition tool

Figure 8. Neural network classification results of the training and testing datasets for breast abnormality detection for PS3 (33 subjects). The results illustrate model tuning and performance evaluation. The image was created using Matlab software.
Abbreviations: CE: Cross-entropy; %E: Percentage of correctly classified elements.

for population-based case–control studies in PS3 and FS, respectively. These matrices offer a visual representation of the classification performance, aiding in assessing the system's accuracy and reliability.

The developed system's exceptional accuracy for screening breast abnormalities and detecting malignant tumors was validated at 93.18%, underscoring its reliability and effectiveness.

Finally, Table 4 provides a comparative analysis of studies conducted in population-based case–control settings, elucidating the progression and refinement of the system across different phases. This comprehensive comparison offers insights into the system's evolution and performance enhancements.

4. Discussion

This system has been installed at a renowned hospital in North-east India, known for its mass screening capabilities. The subsequent product deployment will include installations at various hospitals across India, leveraging the system's superior performance and excellent output based on the second dataset acquired through our proposed IR image acquisition and analysis technique.

The study utilized a double-blind validation method where expert doctors and reviewers provided both quantitative and qualitative feedback. This approach ensured an impartial evaluation of the model's performance, as the experts and reviewers were unaware of the algorithm's predictions during the assessment. The

Figure 9. Neural network classification results of the training and testing datasets for breast abnormality detection for the final study (88 subjects). The results display model tuning and performance evaluation. The image was created using Matlab software.
Abbreviations: CE: Cross-entropy; %E: Percentage of correctly classified elements.

method was rigorously applied across all 88 subjects in the FS phase.

Focused on IR imaging, the study recorded the number of pixels corresponding to each temperature cluster zone and used this information to quantify areas of interest. Given the 640 × 480-pixel IR images, each subject's dataset included seven zones across two ambient temperatures for 16 images (total number of image data = [71 × 2] + [10 × 4] + [33 × 32] + [88 × 32] = 4054).

The study's primary focus was to explore innovative IR imaging techniques and related features for breast cancer screening. We aimed to investigate temperature-based imaging and machine-learning algorithms as alternative diagnostic methods. This approach allowed us to detect thermal patterns and variations that could indicate potential abnormalities, providing a different perspective on breast cancer screening. Although integrating these traditional measures could enhance the study, our concentration was on advancing the field of IR imaging to contribute valuable knowledge to breast cancer screening.

The system's repeatability was confirmed by imaging the same breast 5 times, demonstrating high consistency with minimal variability. Results across trials were consistent, as evidenced by acceptable statistical measures, including standard deviation and coefficient of variation, affirming the system's accuracy and clinical viability for breast cancer screening.

Clinical implications and utility: the findings regarding their clinical implications and the system's utility in breast cancer diagnosis and population screening have been

explored. Results indicate the system's promise as an effective diagnostic tool for early detection of high-risk individuals. The system's non-invasive and non-contact nature makes it well-suited for population screening, despite challenges with ambient temperature adjustments.

After considering the experts' concerns, we conducted a thorough comparative analysis, including using the system with biopsy and USG. This analysis was carried out with great attention to detail. This analysis addresses the expert's and concerned doctors' request for a comparison with established diagnostic methods. The system exhibited a sensitivity comparable to that of biopsy (90.2% vs. 88.6%) and USG (90.2% vs. 89.7%). Furthermore, the system displayed competitive specificity, with respective values of 82.8%, 84.6%, and 82.4% for biopsy, USG, and the system. This comprehensive analysis not only addresses reviewers' concerns but also underscores the potential utility of the system as a valuable diagnostic tool.

5. Conclusion

IR imaging plays a crucial role in various medical applications, emphasizing IR image acquisition techniques. This paper reported different types of IR image acquisition systems based on trials in a hospital setup and conclusively identified the superior one. Key features of the proposed imaging system include a touchless and painless IR camera-based system for maximum patient comfort; gantry rotation for acquiring multiple breast angles; dynamic IR image collection within a temperature-controlled chamber; and a simplified user interface for data collection by technicians, IR image analysis experts, and doctors.

The primary challenge in this study was managing patients with varied health conditions. During the early stages of development, patients had to wait a long time to reach a stable room temperature. However, as the development of the novel data acquisition technique progressed, the evolved system became more user-friendly and efficient regarding imaging quality. The next challenge addressed was maintaining a constant ambient temperature during data collection, which was the most difficult task. It was overcome by implementing a temperature-controllable enclosure. Here, two ambient temperatures have been taken. The higher temperature was 25°C, and the lower temperature was 23°C. The mean temperature adjustment for each subject is 1°C.

Finally, IR image analysis software was developed, incorporating machine learning algorithms that produced excellent results. These findings were cross-validated using USG and biopsy reports. However, several limitations were identified during this study that may have influenced the results and should be considered when interpreting the findings. One potential source of bias arises from the sample population, which may not fully represent the general population due to demographic variations and differences in breast cancer prevalence, thereby potentially limiting the study's generalizability.

Confounding variables such as variations in breast density, tissue composition, and patient positioning during imaging could have affected the accuracy and consistency of the system. Although the study attempted to control for these factors, they may have introduced some degree of variability in the results. External factors such as equipment quality and maintenance, technician expertise, and interpretation differences among medical professionals may have also impacted the study's outcomes. Moreover, relying on USG and biopsy reports to cross-validate the system's performance introduces potential dependencies on the accuracy and reliability of these other diagnostic modalities.

The study progressed iteratively, with each phase's findings and feedback shaping the design and objectives of the subsequent phase. This approach allowed us to refine methods and address challenges progressively. By enhancing techniques and analyses based on results-driven objectives, each phase naturally evolved, focusing on enhancing the efficiency and accuracy of our breast cancer screening imaging system. In terms of sample size, we acknowledge that the varying sizes across phases may impact the overall consistency of the results. However, the number of subjects available for each phase depended on live patient availability during the study period at the hospital. Practical constraints such as time and resource limitations influenced sample sizes, despite efforts to maintain consistency. While we tried to work with consistent sample sizes, external factors such as patient availability and medical considerations posed challenges. However, our phased approach enabled us to optimize methods and techniques, yielding improved results in each subsequent phase. Our study focused on real-world application and practical implementation, requiring flexibility in our approach.

In addition, the challenges faced during data collection, such as maintaining a constant ambient temperature and managing patients with diverse health conditions, affected the precision and consistency of the imaging process. While we addressed these challenges throughout the study, residual variability may have affected the results. Overall, while the study presents promising findings, we carefully considered its limitations and potential sources of bias when evaluating the system's effectiveness and applicability in broader clinical settings. Future research should address

these limitations to further validate and improve the system.

Acknowledgments

We sincerely acknowledge Sayantani Banerjee's technical contributions and Dr. R. Ravi Kannan and his team for providing hospital support for data collection and test beds. In addition, Sri Aditya Kumar Sinha, Center Head, C-DAC, Kolkata, has provided constant support and help with the implementation.

Funding

This project is funded by MeitY (Ministry of Electronics and Information Technology), Government of India, bearing administrative approval No. 1(4)/2015- ME&HI.

Conflict of interest

The authors declare no conflicts of interest.

Author contribution

Conceptualization: Asok Bandyopadhyay, Himanka S. Mondal
Investigation: Himanka S. Mondal, Barnali Pal
Methodology: Himanka S. Mondal
Writing – original draft: Asok Bandyopadhyay, Himanka S. Mondal
Writing – review & editing: Bivas Dam, Dipak C. Patranabis, Asok Bandyopadhyay

Ethics approval and consent to participate

The study conducted in this paper involved human subjects and was approved by the Cachar Cancer Hospital and Research Centre Institutional Review Board (IRB)/ Ethics Committee. The approval number granted by the IRB/Ethics Committee is CCHRC/IRB/01/2019/254. The IRB/Ethics Committee reviewed and approved the study protocol, ensuring compliance with ethical standards and guidelines for research involving human participants. Written informed consent was obtained from each human subject before their participation. The consent form provided detailed information about the study objectives, procedures, potential risks and benefits, confidentiality measures, and the voluntary nature of participation. Subjects were informed of their right to withdraw from the study at any time without consequences. Consent forms were securely stored in compliance with data protection regulations.

Consent for publication

Participants gave consent to publish their data in the consent paper.

Availability of data

The data utilized in this study are proprietary to the Ministry of Electronics and Information Technology (MeitY), Government of India. Therefore, it cannot be disclosed due to confidentiality agreements and is not available for external access or disclosure.

References

1. Sivanandam S, Anburajan M, Venkatraman B, Menaka M, Sharath D. Medical thermography: A diagnostic approach for type 2 diabetes based on non-contact infrared thermal imaging. *Endocrine*. 2012;42(2):343-351.

   doi: 10.1007/s12020-012-9645-8

2. Nishide K, Nagase T, Oba M, *et al.* Ultrasonographic and thermographic screening for latent inflammation in diabetic foot callus. *Diabetes Res Clin Pract*. 2009;85(3):304-309.

   doi: 10.1016/j.diabres.2009.05.018

3. Fujiwara Y, Inukai T, Aso Y, Takemura Y. Thermographic measurement of skin temperature recovery time of extremities in patients with type 2 diabetes mellitus. *Exp Clin Endocrinol Diabetes.* 2000;108(7):463-469.

   doi: 10.1055/s-2000-8142

4. Fushimi H, Inoue T, Yamada Y, Matsuyama Y, Kubo M, Kameyama M. Abnormal vasoreaction of peripheral arteries to cold stimulus of both hands in diabetics. *Diabetes Res Clin Pract*. 1996;32(1-2):55-59.

   doi: 10.1016/0168-8227(96)01222-3

5. Sodi A, Giambene B, Miranda P, Falaschi G, Corvi A, Menchini U. Ocular surface temperature in diabetic retinopathy: A pilot study by infrared thermography. *Eur J Ophthalmol*. 2009;19(6):1004-1008.

   doi: 10.1177/112067210901900617

6. Ring EFJ, Ammer K. Infrared thermal imaging in medicine. *Physiol Meas*. 2012;33(3):R33-R46.

   doi: 10.1088/0967-3334/33/3/R33

7. EtehadTavakol M, Sadri S, Ng EYK. Application of K- and Fuzzy c-means for color segmentation of thermal infrared breast images. *J Med Syst*. 2010;34(1):35-42.

   doi: 10.1007/s10916-008-9213-1

8. Etehadtavakol M, Ng EYK. *Color Segmentation of Breast Thermograms: A Comparative Study*. Singapore: Springer; 2017. p. 69-77.

   doi: 10.1007/978-981-10-3147-2_6

9. Pramanik S, Banik D, Bhattacharjee D, Nasipuri M, Bhowmik MK, Majumdar G. Suspicious-region segmentation from breast thermogram using DLPE-based level set method. *IEEE Trans Med Imaging.* 2019;38(2):572-584.

   doi: 10.1109/TMI.2018.2867620

10. EtehadTavakol M, Chandran V, Ng EYK, Kafieh R. Breast cancer detection from thermal images using bispectral invariant features. *Int J Therm Sci.* 2013;69:21-36.

doi: 10.1016/j.ijthermalsci.2013.03.001

11. Garduño-Ramón MA, Vega-Mancilla SG, Morales-Henández LA, Osornio-Rios RA. Supportive noninvasive tool for the diagnosis of breast cancer using a thermographic camera as sensor. *Sensors* (*Basel*). 2017;17(3):497.

doi: 10.3390/s17030497

12. Prakash RM, Bhuvaneshwari K, Divya M, Sri KJ, Begum AS. Segmentation of Thermal Infrared Breast Images Using K-means, FCM and EM Algorithms for Breast Cancer Detection. In: *2017 International Conference on Innovations in Information, Embedded and Communication Systems* (*ICIIECS*), IEEE; 2017. p. 1-4.

doi: 10.1109/ICIIECS.2017.8276142

13. Venkataramani K, Mestha LK, Ramachandra L, Prasad SS, Kumar V, Raja PJ. Semi-automated breast cancer tumor detection with thermographic video imaging. *Annu Int Conf IEEE Eng Med Biol Soc.* 2015;2015:2022-2025.

doi: 10.1109/EMBC.2015.7318783

14. Kandlikar SG, Perez-Raya I, Raghupathi PA, *et al*. Infrared imaging technology for breast cancer detection - Current status, protocols and new directions. *Int J Heat Mass Transf*. 2017;108:2303-2320.

doi: 10.1016/j.ijheatmasstransfer.2017.01.086

15. Bandyopadhyay A, Chaudhuri A, Mondal HS. IR based Intelligent Image Processing Techniques for Medical Applications. In: *2016 SAI Computing Conference* (*SAI*), IEEE; 2016. p. 113-117.

doi: 10.1109/SAI.2016.7555970

16. Bandyopadhyay A, Mondal HS, Dam B, Patranabis DC. Efficient infrared image processing and machine learning algorithm for breast cancer screening. *Comput Methods Biomech Biomed Eng Imaging Vis*. 2023;11:2226-2238.

doi: 10.1080/21681163.2023.2225639

17. Bandyopadhyay A, Mondal HS, Pal B, Dam B, Patranabis DC. Exploring the Potential Use of Infrared Imaging in Medical Diagnosis: A Comprehensive Framework for Diabetes and Breast Cancer Screening. In: *Proceedings of 4th International Conference on Image Processing and Capsule Networks* (*ICIPCN*), Springer; 2023.

18. Kapoor P, Prasad SVA. Image Processing for Early Diagnosis of Breast Cancer Using Infrared Images. In: *2010 the 2nd International Conference on Computer and Automation Engineering* (*ICCAE*), IEEE; 2010. p. 564-566.

doi: 10.1109/ICCAE.2010.5451827

19. Gonzalez RC, Woods RE. *Digital Image Processing*. 4th ed. London: Pearson Education, Inc.; 2007.

20. Chen GL, Lee CY. Iterative Morphology-based Segmentation of Breast Tumors in Ultrasound Images. In: *2014 International Symposium on Computer, Consumer and Control*, IEEE; 2014. p. 1107-1110.

doi: 10.1109/IS3C.2014.288

21. Li C, Xu C, Gui C, Fox MD. Distance regularized level set evolution and its application to image segmentation. *IEEE Trans Image Process.* 2010;19(12):3243-3254.

doi: 10.1109/TIP.2010.2069690

22. Caselles V, Catté F, Coll T, Dibos F. A geometric model for active contours in image processing. *Numer Math*. 1993;66:1-3.

doi: 10.1007/BF01385685

23. Mambou SJ, Maresova P, Krejcar O, Selamat A, Kuca K. Breast cancer detection using infrared thermal imaging and a deep learning model. *Sensors* (*Basel*). 2018;18(9):2799.

doi: 10.3390/s18092799

24. Tsietso D, Yahya A, Samikannu R. A review on thermal imaging-based breast cancer detection using deep learning. *Mob Inf Syst.* 2022;2022:1-19.

doi: 10.1155/2022/8952849

25. Wu MN, Lin CC, Chang CC. Brain Tumor Detection Using Color-Based K-Means Clustering Segmentation. In: *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (*IIH-MSP*), IEEE; 2007. p. 245-250.

doi: 10.1109/IIHMSP.2007.4457697

26. Zhang Y, Wu X, He L, *et al*. Applications of hyperspectral imaging in the detection and diagnosis of solid tumours. *Transl Cancer Res*. 2020;9(2):1265-1277.

doi: 10.21037/tcr.2019.12.53

27. Lugano R, Ramachandran M, Dimberg A. Tumor angiogenesis: Causes, consequences, challenges and opportunities. *Cell Mol Life Sci.* 2020;77(9):1745-1770.

doi: 10.1007/s00018-019-03351-7

28. Houssein EH, Emam MM, Ali AA. An efficient multilevel thresholding segmentation method for thermography breast cancer imaging based on improved chimp optimization algorithm. *Expert Syst Appl*. 2021;185:115651.

doi: 10.1016/j.eswa.2021.115651

29. Houssein EH, Abdelkareem DA, Emam MM, Hameed MA, Younan M. An efficient image segmentation method for skin cancer imaging using improved golden jackal optimization algorithm. *Comput Biol Med*. 2022;149:106075.

doi: 10.1016/j.compbiomed.2022.106075

30. De Santana MA, Pereira JMS, da Silva FL, *et al.* Breast cancer diagnosis based on mammary thermography and extreme learning machines. *Res Biomed Eng*. 2018;34(1):45-53.

doi: 10.1590/2446-4740.05217

31. Gonçalves CB, Leles ACQ, Oliveira LE, Guimaraes G, Cunha JR, Fernandes H. Machine learning and infrared thermography for breast cancer detection. *Proceedings*. 2019;27:45.

doi: 10.3390/proceedings2019027045

32. Dixit S, Kumar A, Srinivasan K. A current review of machine learning and deep learning models in oral cancer diagnosis: Recent technologies, open challenges, and future research directions. *Diagnostics* (*Basel*). 2023;13(7):1353.

doi: 10.3390/diagnostics13071353

33. Cohen EE, Ahmed O, Kocherginsky M, *et al.* Study of functional infrared imaging for early detection of mucositis in locally advanced head and neck cancer treated with chemoradiotherapy. *Oral Oncol.* 2013;49(10):1025-1031.

doi: 10.1016/j.oraloncology.2013.07.009

34. *Advanced Thermography and Preventive Education*. Available from: https://thermogramcenter.com [Last accessed on 2023 Dec].

35. Ng EYK. A review of thermography as promising non-invasive detection modality for breast tumor. *Int J Therm Sci.* 2009;48(5):849-859.

doi: 10.1016/j.ijthermalsci.2008.06.015

36. Wishart GC, Campisi M, Boswell M, *et al.* The accuracy of digital infrared imaging for breast cancer detection in women undergoing breast biopsy. *Eur J Surg Oncol.* 2010;36(6):535-540.

doi: 10.1016/j.ejso.2010.04.003

37. Kakileti ST, Madhu H, Subramoni T, Manjunath G. Thermalytix: Using AI to save lives. *XRDS Crossroads ACM Mag Stud.* 2020;26(3):38-41.

doi: 10.1145/3383384

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Dental cavity analysis, prediction, localization, and quantification using computer vision

Mohammad Aqeel[1], Payam Norouzzadeh[2], Abbas Maazallahi[1], Salih Tutun[3], Golnesa Rouie Miab[4], Laila Al Dehailan[5], David Stoeckel[6], Eli Snir[3], and Bahareh Rahmani[1]*

[1]Computer Science, Saint Louis University, St. Louis, Missouri, United States of America

[2]Professional Studies, Saint Louis University, St. Louis, Missouri, United States of America

[3]Data Analytics Area, Olin Business School, Washington University in Saint Louis, St. Louis, Missouri, United States of America

[4]Pacific Dental Services, St. Louis, Missouri, United States of America

[5]Department of Restorative Dental Sciences, College of Dentistry, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia

[6]Department of Dentistry, Saint Louis University, St. Louis, Missouri, United States of America

Abstract

Dental health assessment is a critical component of overall well-being, and advancements in computer vision and deep learning have opened new avenues for automating and enhancing this process. In this study, we present a comprehensive approach to dental cavity analysis, spanning localization, quantification, and visualization. Our methodology leveraged a diverse dataset of colored dental images that had been meticulously augmented and annotated. The You Only Look Once model was employed for precise dental cavity localization, providing bounding box predictions. Remarkably, these results were obtained based on images from standard device cameras. Subsequently, we introduced the use of the segment anything model segmentation model, known for its zero-shot generalization capabilities, to focus on the exact areas of dental cavities. This approach enhanced the granularity of our analysis, providing dental professionals with detailed visualizations for precise diagnosis. During the quantification phase, we extracted cavity areas from bounding box coordinates, enabling accurate measurement of cavity sizes. The model achieved a notable mean average precision of 0.732, an accuracy of 0.789, and a recall of 0.701. Moreover, the model converged quickly, with most metrics achieving near-optimal results after 100 iterations. This quantitative data augments traditional diagnosis methods, facilitating more informed treatment decisions.

Keywords: You only look once; Segment anything model; Segmentation model; Dental cavity

***Corresponding author:**
Bahareh Rahmani
(brahmani@slu.edu)

Citation: Aqeel M, Norouzzadeh P, Maazallahi A, *et al*. Dental cavity analysis, prediction, localization, and quantification using computer vision. *Artif Intell Health*. 2024;1(3):80-88.
doi: 10.36922/aih.3184

Received: March 15, 2024

Accepted: May 14, 2024

Published Online: July 24, 2024

1. Introduction

Many may not be aware that our oral well-being can provide insights into our overall health. The truth is that problems in our mouth can potentially impact the rest of our body. Similar to other parts of our body, our mouths harbor mostly harmless bacteria.

However, some of these bacteria have the potential to cause health problems, as they serve as a gateway to our digestive and respiratory systems.

The body's own defense typically keeps these germs in check, along with basic dental hygiene habits like frequent brushing and flossing. However, without proper oral hygiene, bacteria levels can rise and cause ailments such as tooth decay and gum disease. Research indicates that certain medical conditions may be impacted by oral bacteria and the inflammation brought on by chronic gum disease, also known as periodontitis. Our oral health might contribute to various diseases and conditions, including endocarditis, cardiovascular disease, and pneumonia.[1]

According to the Global Burden studies in 2019, dental caries is the most common oral disease, affecting around 3.5 billion people, of whom 2 billion have permanent dental caries.[2] Moreover, 1.45 million of the 6 million patients with dental caries who visited dentists in the Republic of Korea in 2020 were children (0 – 9 years old).[3] Therefore, it is important to study tooth caries.

Caries formation is affected by a host of preferential habits, systemic disorders, and congenital anomalies. Incipient carious lesions are frequently overlooked by patients. Dentists treat them conservatively using techniques of minimally invasive dentistry. As a result, there are many instances of misdiagnosis and poor care, particularly among young practitioners conducting visual and radiographic investigations. Caries mismanagement can be expensive and leave the patient exposed to future periapical, osseous, and fascial space spread of the infection.[4] These challenges are particularly noteworthy in areas with limited access to advanced dental facilities and trained practitioners. Early detection and ongoing monitoring of these problems are proposed in this article using a low-cost automated system that does not differentiate patients based on their sociodemographic status.

Since cellular technology has expanded globally, even to rural areas, the use of mobile portable devices like smartphones has increased exponentially in emerging economies.[5] As a result, biomedical research can harness functionality in smartphones to offer cost-effective solutions to challenging issues in oral treatment.

In this research endeavor, we define several core objectives, each serving as a pivotal milestone in our pursuit of advancing dental cavity analysis through computer vision techniques. These objectives collectively constitute the foundation upon which our study is constructed, steering our research toward meaningful and innovative contributions to the field. Our first and foremost objective is to pioneer the development of a robust computer vision model dedicated to the precise localized identification of dental cavities within colored images. We recognize that the accurate identification of cavity locations is an indispensable initial step in streamlining the diagnostic process. Achieving this objective will empower dental professionals with a powerful tool that not only identifies cavities but also provides their exact spatial coordinates.

Building upon the success of our initial objective, our second key goal is to introduce a methodology for the quantification of dental cavity areas within these localized regions. Accurate area quantification is paramount in assessing the severity of cavities and can significantly aid in treatment planning. With this objective, we aim to automate and standardize the area measurement process, thereby enhancing the precision of dental health assessments. Furthermore, it is noteworthy to establish a subtle connection between our current study's objectives and our prior research endeavors. In our earlier work, we endeavored to forecast dental cavities utilizing convolutional neural networks (CNNs). While that research focused on predicting the occurrence of cavities, the objectives of the current study extend beyond prediction. Here, we embark on the critical task of geospatial mapping them within images, quantifying their extent, and subsequently facilitating a more comprehensive understanding of their impact on oral health. This intrinsic linkage between our research pursuits underscores the holistic approach we adopt in addressing the multifaceted challenges in dental cavity analysis, paving the way for a more comprehensive and integrated solution.

In this study, we collected a dataset of colored images containing dental cavities and manually annotated this dataset using Roboflow annotation. We then trained the "You Only Look Once" (YOLO) v5 model to detect and locate dental cavities in these images using a bounding box. Once we identified the exact cavity area with a bounding box, we used the coordinates of the bounding box to calculate the area of the cavity. Applying image segmentation on the cavity highlighted the cavity area, and cavity masks were obtained from segmentation for further analysis of the dental cavity shape.

2. Literature review

In our prior research endeavor, entitled "Forecasting teeth cavities by CNNs," we conducted a comprehensive exploration into predicting dental cavities using CNNs. The dataset in this previous investigation consisted of X-ray images. To augment the dataset's size and diversify its content, we applied a series of sophisticated augmentation techniques. To further enhance the accuracy and efficacy of dental cavity prediction, we methodically incorporated

segmentation techniques into the research framework. Four distinct segmentation methods were evaluated, namely segmentation with the thresholding method, segmentation with the contouring method, segmentation with the Canny-edges method, and segmentation with a combination of these techniques. The best performance among all methods was obtained by the Canny edge-CNN mode.[6]

In response to the inefficiency and complexity of traditional dental disease detection methods, a study introduced a novel approach utilizing deep learning.[7] The study employed the YOLOv3 model to automate the detection and classification of four common teeth problems: cavities, root canals, dental crowns, and broken-down root canals, using panoramic dental X-ray images orthopantograms. To overcome data limitations, a dental X-ray dataset with 1200 augmented images was created and divided into 70% for training and 30% for testing. The YOLOv3 model achieved a remarkable 99.33% accuracy, outperforming existing models and demonstrating its versatility with other datasets.[7]

Another study indicated that deep learning models can be used to help dentists in planning dental implant placement, ensuring that dental implants are optimally placed and properly aligned with the surrounding teeth and bone.[8]

Dental caries, one of the most prevalent dental conditions in contemporary times, poses significant challenges for early detection in dental X-ray or radiovisiography images. Deep learning has been widely employed across medical domains for predictive and diagnostic purposes. One of the investigations evaluated a K-means clustering approach for image segmentation, underscoring the significance of image enhancement techniques in improving the quality of dental radiographs. The implemented K-means model algorithm demonstrated improved accuracy in the detection of dental caries.[9]

Tareq *et al.*,[10] aimed to pioneer a novel and cost-effective virtual computer vision artificial intelligence (AI) system capable of predicting dental cavitation from non-standardized photographs with reasonable clinical accuracy. They curated a dataset comprising 1703 augmented images sourced from 233 de-identified teeth specimens, captured using consumer-grade smartphones. The methodology leveraged cutting-edge techniques, including ensemble modeling, test-time augmentation, and transfer learning processes. The researchers independently assessed derivatives of the YOLO algorithm, including v5s, v5m, v5l, and v5x, subsequently creating an ensemble model and transfer-learning it with ResNet50, ResNet101, VGG16, AlexNet, and DenseNet. Evaluation metrics encompass precision, recall, and mean average precision (mAP). The YOLO model ensemble achieved a notable mAP of 0.732, an accuracy of 0.789, and a recall of 0.701. When applied to VGG16, the final model demonstrated impressive diagnostic accuracy of 86.96%, with precision and recall values of 0.89 and 0.88, respectively. This performance outstripped all other existing methods for object detection in free-hand, non-standardized smartphone photographs. The virtual computer vision AI system, enriched by an ensemble model, test-time augmentation, and transfer learning techniques, successfully predicts dental cavitations from non-standardized photographs with clinically reasonable accuracy. This innovation holds the potential to enhance access to oral health care in resource-constrained, underserved areas and facilitates automated diagnostics and advanced tele-dentistry applications.

Thanh *et al.*,[11] demonstrated the potential of mobile phone-based diagnostic tools for dental caries detection using deep learning algorithms, highlighting the efficiency of YOLOv3 and Faster R-CNN models. A blog article on innovative applications in dentistry[12] showcased AIs ability to detect caries with high accuracy using image augmentation and transfer learning, emphasizing its role in complementing traditional diagnostic methods. In addition, a GitHub project has been established, aiming to detect and localize various dental diseases, including caries and periodontal diseases, using computer vision in panoramic dental X-ray images.[13]

Nakai and Wei,[14] while focusing on protein localization, highlighted the adaptability of deep learning techniques, such as CNN and long short-term memory, for predictive modeling across diverse fields, including dentistry. Acharya[15] discussed deep learning techniques for image segmentation, including U-Net and SegNet, which are crucial for detailed analysis in medical imaging and diagnostics. Brownlee[16] explored the architectures of Fast R-CNN and Faster R-CNN for real-time object detection, relevant for precise localization and quantification in dental imaging. Fernandes *et al.*,[17] while focused on animal sciences, underscored the importance of machine learning and deep learning algorithms in various computer vision applications, illustrating the multidisciplinary potential of these technologies.

A study on gait pattern recognition for flat fall prediction highlighted the use of computer vision and machine learning in recognizing gait patterns, demonstrating the versatility of these technologies in health diagnostics beyond dental applications.[18] A notable study utilized CNNs to diagnose dental caries from bitewing images, emphasizing the complexity of identifying proximal and interproximal dental caries and the effectiveness

of bitewing images in clearly capturing such lesions.[19] Another innovative approach involved classifying tooth caries using quantitative light-induced fluorescence (QLF) images with the help of the Xception deep learning model, underscoring the significance of image augmentation and K-fold cross-validation in training robust models.[20] A systematic review aimed at evaluating neural networks in caries detection highlighted the diverse methodologies and neural network architectures employed across studies, reflecting the dynamic evolution of AI applications in dental diagnostics.[21]

Further illustrating the potential of machine learning in dentistry, a previous study applied several algorithms, notably random forest, achieving high performance in predicting the risk of dental caries from a dataset derived from a children's oral health survey.[22] A systematic review focusing on AI for radiographic imaging detection of caries lesions critically evaluated studies, revealing a preference for CNN models in most research, with a range from 15 to 2900 radiographs used across various studies to build AI models.[23] The use of deep learning for caries detection through tooth surface segmentation in intraoral photographic images has been investigated, employing U-Net for segmentation and ResNet-18 and Faster R-CNN for classification and localization, thereby reducing false alarms and enhancing detection accuracy.[24] Another study developed a CNN model for diagnosing dental caries from bitewing radiographs, demonstrating the utility of deep learning in enhancing dental diagnostic processes.[19]

A research endeavor introduced a novel method for classifying dental caries using QLF imaging combined with CNNs, aiming to improve accuracy in real-time caries detection in clinical settings.[20] Lian *et al.*,[25] utilized deep learning methods to detect and classify caries lesions on panoramic films, comparing performance with expert dentists and showing similar accuracy and reliability. Alharbi *et al.*[26] applied nested U-Net models to dental panoramic X-ray images for caries detection, demonstrating high testing accuracy and robust model performance.

Sikri *et al.*,[27] presented a comprehensive narrative review on the applications of AI in dentistry, detailing how AI integrates into various aspects of dental care, from diagnostics to patient management. Meanwhile, Zhou *et al.*[28] explore a more focused application with their development of a context-aware CNN specifically designed for diagnosing caries in children from dental panoramic radiographs, demonstrating the potential of machine learning to address unique challenges in pediatric dentistry.

These studies collectively underscore the transformative impact of machine learning and AI on dental diagnostics, heralding a new era of precision and efficiency in detecting dental caries. As we delve further into this article, we will explore the mechanics behind these innovations, their practical applications, and the challenges and future directions in integrating advanced computational techniques into dental care.

3. Methods

3.1. Data collection and pre-processing

The image data were sourced from Kaggle (https://www.kaggle.com/datasets/salmansajid05/oral-diseases?resource=download-directory). This dataset comprises a collection of images obtained from multiple health centers and reliable dental websites, ensuring the variety and validity of the dental conditions depicted. Each image in the dataset is thoroughly marked with bounding boxes, accurately representing the dental condition.

3.1.1. Description of the colored image dataset

The colored image dataset used in this study comprises a total of 218 dental cavity images captured using a standard device camera. These images were obtained under casual conditions, featuring open jaws and clear representations of dental cavities. The dataset served as the foundational source of visual data for training and evaluation (Figure 1).

3.1.2. Data augmentation techniques

Data augmentation plays a pivotal role in expanding the dataset and enhancing model robustness. To achieve this, we leveraged the Image Data Generator, an image augmentation API integrated within Keras – an open-source Python library for machine learning. ImageDataGenerator enabled artificially diversifying the dataset by applying transformations such as rotation, shifting, zooming, shearing, and reflection. These augmentations fostered the development of more adept models and improved their ability to generalize across various scenarios. In our experimentation, the augmentation parameters were set as follows:

i. Rotation range: 40°
ii. Width and shifting range: 0.2
iii. Zoom range: 0.2
iv. Shear range: 0.2.

Figure 1. Colored image with a single cavity and multiple cavities

Applying these augmentation techniques expanded the dataset to a total of 2383 images, thereby facilitating a richer and more comprehensive training process.

3.1.3. Manual annotation process

Following the augmentation phase, we proceeded with the manual annotation of dental cavities within the augmented images. For this purpose, we employed the Roboflow annotation tool, which facilitated meticulous annotation of dental cavities (Figure 2). The chosen method for object detection involved bounding box annotation, represented by a rectangular box icon within the annotation tool. In the annotation process, annotators utilized crosshairs to determine the starting point for drawing bounding boxes around dental cavities. Each bounding box served as an annotation for the presence and location of a dental cavity within the image. Furthermore, the Class Selector within the tool allowed annotators to assign the appropriate label to each annotated bounding box, signifying the presence of a dental cavity. This manual annotation process was performed for approximately 400 images, ensuring the availability of accurately labeled data for the subsequent training of the object detection model focused on dental cavity identification.

3.2. Dental cavity localization using YOLOv5

The schematic of the research project is shown in Figure 3A. Once the predictive model is developed, its application is straightforward. Implementation can be developed for a smartphone app, where images would be taken by patients or dental assistants, without the need for a dental professional. Figure 3B describes how the prescriptive model would be used.

3.2.1. Introduction to YOLO

In our pursuit of precise dental cavity localization with the augmented dataset, we harnessed the power of the YOLO object detection framework. YOLO represents a groundbreaking approach to object detection, characterized by its remarkable speed and accuracy. Unlike traditional object detection models, YOLO processes images in a single pass, making it especially efficient for real-time applications. The YOLO algorithm divides an image into a grid and predicts bounding boxes and associated class probabilities for each grid cell. This unique methodology enables YOLO to excel in scenarios where objects of interest may vary in size and scale, making it precisely suited for our task of dental cavity localization.

3.2.2. Training YOLOv5 model

The YOLOv5 model, an evolution of the YOLO architecture, served as the cornerstone of our dental cavity localization efforts. Training the YOLOv5 model involved

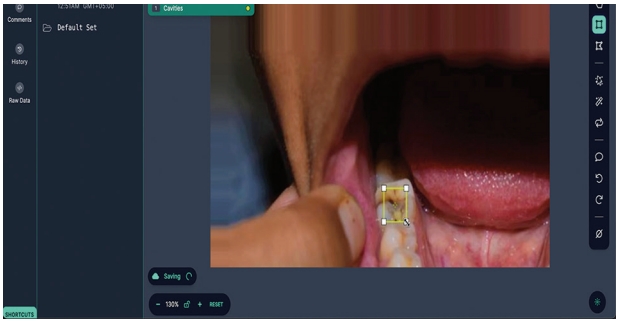


Figure 2. Annotation using the Roboflow annotation tool

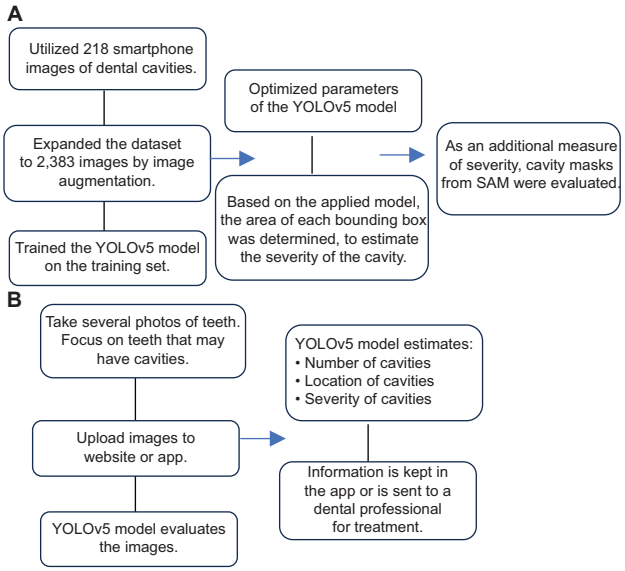


Figure 3. (A) The schematic of the research project. Training set will be trained and optimized by YOLOv5 Model. SAM (Segment Anything Model) will be evaluated. (B) The use of the prescriptive model. The number of cavities, locations and severity of them will be estimated. Abbreviation: SAM: Segment anything model.

a systematic process aimed at enabling it to accurately predict the presence and location of dental cavities within our annotated images. Our training dataset, enriched by augmentation techniques and manual annotations, was used to train the YOLOv5 model. This training process involved iteratively fine-tuning the model's parameters and optimizing its ability to recognize dental cavities in varying image contexts. The YOLOv5 model's training process was rigorous, ensuring a high degree of accuracy and robustness in detecting dental cavities within the images. The successful training of this model constituted a crucial milestone in our endeavor to automate dental cavity localization, facilitating more precise and efficient dental health assessments.

3.2.3. Bounding box prediction

Following the successful training of the YOLOv5 model, we transitioned to the crucial phase of bounding box

prediction and post-processing. This step represents the culmination of our efforts to precisely locate dental cavities within unknown images, a process that significantly contributes to the automation of dental health assessment. The YOLOv5 model, trained on our annotated dataset, acquired the capability to predict bounding boxes around dental cavities with remarkable accuracy. To employ this predictive power, we utilized a streamlined command that swiftly and accurately delineates the region of dental cavities when applied to an unknown image. These bounding boxes serve as visual indicators of cavity presence and location within the image (Figure 4).

3.3. Quantification of cavity area

3.3.1. Extracting cavity area from bounding box

In our pursuit of a comprehensive dental cavity analysis, the localization of cavities through bounding box predictions facilitated by the YOLOv5 model marked a significant milestone. With these bounding boxes accurately delineating the regions of interest, the next logical step in our research was to quantify the area encompassed by these bounding boxes, effectively measuring the extent of dental cavities in pixels.

The extraction of cavity area from the bounding boxes generated by the YOLOv5 model is a straightforward yet essential process. The model's coordinates, specifically (Xmin, Ymin, Xmax, Ymax), facilitate straightforward calculation of the area of the contained bounding region.

The following is a brief breakdown of the steps involved (Figure 5):

i. Width calculation (width): We subtract the Xmin coordinate from the Xmax coordinate, where the result represents the horizontal span of the cavity region, to determine the width of the bounding box.

$$Width = Xmax - Xmin \quad (I)$$

ii. Height calculation (height): To represent the vertical extent of the cavity region, we calculate the difference between the Ymin coordinate and the Ymax coordinate.

$$Height = Ymax - Ymin \quad (II)$$

iii. Area computation (area): The final step involves calculating the area of the cavity region by multiplying the width by the height, yielding the area in pixels.

$$Area = width \times height \quad (III)$$

By systematically employing these calculations, we can precisely quantify the area of each dental cavity within the images. This precise quantification empowers dental professionals with valuable information for assessing cavity severity and planning appropriate treatment interventions.

Figure 4. Single cavity (left panel) and multiple cavities (right panel) detection using the Yolo V5 model

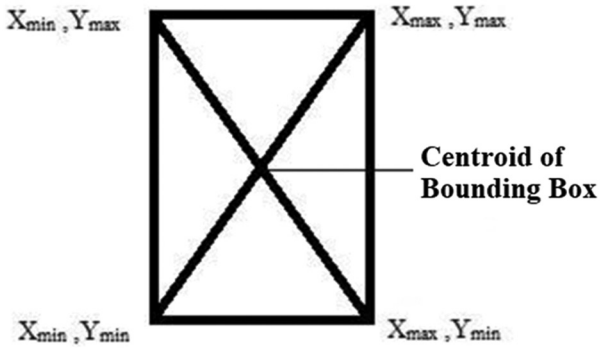


Figure 5. Centroid of a bounding box

4. Results and discussion

4.1. YOLOv5 results and limitations

The YOLOv5 algorithm effectively identified cavities through the bounding box process. The algorithm converges quite quickly, enabling implementation in various applications. As expected, object loss in the training set continuously improves with iterations of the algorithm. However, based on the validation set, overfitting starts to become evident after 100 iterations. Other metrics, such as precision, recall, and mAP, converge after 100 trials, indicating that 100 trials are sufficient and desirable to train the algorithm.

From the results, both box loss and object loss are below 0.03 after 100 iterations. The algorithm achieves an accuracy of 0.789, a recall of 0.701, and an mAP of 0.732. These results validate that pictures from smartphones can be an effective 1st step in identifying and treating dental cavities.

There are several limitations in this study, both in terms of the data collected and the modeling. Camera images can only identify cavities that have already formed on the surface of teeth. In addition, it may be difficult to take images within the mouth. To identify issues that are not easily visible, dental X-rays are required. These challenges are inevitable in any visual technique.

The data in this study are based on image augmentation imitating multiple possible alternatives for each original image. While we believe that this is an accurate representation of future images that will be available, it would be more effective to have actual images from multiple angles for teeth and cavities.

More broadly, as AI is applied to more diverse opportunities in modeling and medical diagnostics, several issues may emerge. These relate both to the development of new models and the use of automated diagnostics by individuals and medical professionals. On the modeling side, one can foresee, in the not-too-distant future, the possibility of automated modeling being employed, evaluating a broad set of models on a given dataset. Without supervision and effective parameter tuning, these methods could lead to overfitting or the use of inappropriate models. Similarly, the data used for these automated studies could be suspect. Using available images that are not evaluated by people could be unreliable. Imagine a situation where an autonomous model is developed by images created by AI, for example.

The implications of AI on medical practice should also be considered. Applications like the one proposed here provide effective but limited self-diagnosing opportunities to individuals, especially in areas with limited access to health. However, it is likely that some people who could receive effective diagnoses from medical professionals would also use these tools. Given the noticeable rate of false negative results, these individuals may not receive the necessary treatment. There are also implications for medical professionals. As medical professionals become more reliant on technology, there is a risk of decreased expertise in the profession. This decline in expertise may arise from becoming overly reliant on forecasting tools or from outsourcing diagnostics to low-cost services that rely on technology.

4.2. Image segmentation using the segment anything model (SAM)

The SAM is a promptable segmentation system capable of zero-shot generalization to unfamiliar objects and images without the need for additional training. This capability allows SAM to segment objects into new images; it has never seen before simply by providing it with a prompt such as a text description, a bounding box, or a few clicks on the image.

SAM is trained on a massive dataset of over 1 billion segmentation masks, making it the largest segmentation dataset to date. This extensive training allows SAM to learn a wide range of object appearances and relationships, enabling it to generalize to new images with high accuracy.

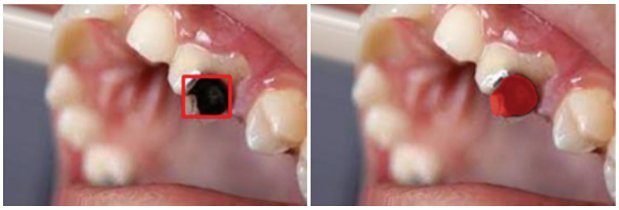


Figure 6. Source image and segmented image

Figure 7. Mask image of the cavity

The SAM is also highly efficient, making it suitable for real-time applications. It can generate a segmentation mask for any prompt in real time after precomputing the image embedding.

For instance, given a bounding box around a dental cavity in an image, the SAM can be used to segment only that area using the SamPredictor class. The SamPredictor class takes a bounding box as input and outputs a segmentation mask for the cavity that is enclosed by the bounding box (Figure 6).

4.3. Dental cavity masks from SAM

Image masks are binary images that represent the foreground pixels of an object. These masks were used to analyze the shape of the dental cavities inside the bounding box. Using the SAM, a segmentation mask for the cavities was generated. The segmentation mask was then cropped to the bounding box and analyzed to measure the desired properties of the cavities. For example, the area of the cavities can be measured by counting the number of white pixels in the cropped segmentation mask.

The white pixels in the binary-masked image show the exact shape of the cavity (Figure 7).

5. Conclusion

In this study, we embarked on a journey to revolutionize dental cavity analysis, resulting in a holistic framework that

redefines the way we diagnose and assess oral health. The YOLO model ensemble achieved a notable mAP of 0.732, an accuracy of 0.789, and a recall of 0.701. Considering that this method identifies cavities directly from standard device camera photographs, this accuracy is remarkable. Our approach, comprising precise localization, accurate quantification, and nuanced visualization, demonstrates its potential to improve dental health assessments to unprecedented levels of accuracy and efficiency. Through meticulous augmentation and annotation of a colored dental image dataset, we harnessed the power of the YOLOv5 model for dental cavity localization, providing bounding box predictions with remarkable accuracy. The introduction of the SAM brought the ability to focus with surgical precision on dental cavities, enriching our analysis and empowering dental professionals with detailed visualizations for diagnosis. Our innovative quantification methodology, which extracts cavity areas from bounding box coordinates, offers a quantitative edge, enhancing diagnostic insights. The potential of our research is promising. We foresee the continued evolution of dental health assessments through the fusion of technology and health care. In the future, our framework could pave the way for automated dental check-ups, reducing the burden on both patients and health-care professionals. Moreover, as technology advances, we envision our methods becoming even more accurate and efficient. Remote dental diagnostics, tele-dentistry, and improved oral health-care access may become more widespread, particularly in underserved areas.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Mohammad Aqeel, Payam Norouzzadeh, Abbas Maazallahi, Eli Snir, Bahareh Rahmani

Investigation: Mohammad Aqeel, Golnesa Rouie Miab, Laila Al Dehailan, David Stoeckel, Bahareh Rahmani

Methodology: Mohammad Aqeel, Payam Norouzzadeh, Salih Tutun, Eli Snir, Bahareh Rahmani

Writing – original draft: Mohammad Aqeel

Writing – review & editing: Payam Norouzzadeh, Abbas Maazallahi, Salih Tutun, David Stoeckel, Eli Snir, Bahareh Rahmani

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The image data can be obtained from Kaggle (https://www.kaggle.com/datasets/salmansajid05/oral-diseases?resource=download-directory).

References

1. Pruthi S. *A Window to Your Overall Health, Oral Health*. United States: Mayo Clinic; 2024.
2. Global Burden of Disease (GBD). Institute for Health Metrics and Evaluation, IHME; 2019.
3. Kim SM, Jang WM, Ahn HA, Park HJ, Ahn HS. Korean National Health Insurance value incentive program: Achievements and future directions. *J Prev Med Public Health*. 2012;45:148-155.

   doi: 10.3961/jpmph.2012.45.3.148
4. Maru AM, Narendran S. Epidemiology of dental caries among adults in a rural area in India. *J Contemp Dent Pract*. 2012;13(3):382-388.

   doi: 10.5005/jp-journals-10024-1155
5. Shankar V, Narang U. Emerging market innovations: Unique and differential drivers, practitioner implications, and research agenda. *J Acad Mark Sci*. 2020;48:1030-1052.

   doi: 10.1007/s11747-019-00685-3
6. Silvertown JD, Wong BP, Abrams SH, Sivagurunathan KS, Mathews SM, Amaechi BT. Comparison of the canary system and DIAGNOdent for the *in vitro* detection of caries under opaque dental sealants. *J Investig Clin Dent*. 2017;8(4).

   doi: 10.1111/jicd.12239
7. Almalki YE, Imam Din A, Ramzan M, *et al*. Deep learning models for classification of dental diseases using orthopantomography X-ray OPG images. *Sensors* (*Basel*). 2022;22(19):7370.

   doi: 10.3390/s22197370
8. Retrouvey JM, Conley RS. Decoding deep learning applications for diagnosis and treatment planning. *Dent Press J Orthod*. 2023;27.

   doi: 10.1590/2177-6709.27.5.e22spe5
9. Kumar S, Kumar H. Analysis of image segmentation techniques for dental radiography. *Element Educ Online*. 2021;20(4):3868-3875.

   doi: 10.17051/ilkonline.2021.04.422
10. Tareq A, Faisal MI, Islam S, *et al*. Visual diagnostics of

dental caries through deep learning of non-standardised photographs using a hybrid YOLO ensemble and transfer learning model. *Int J Environ Res Public Health*. 2023;20:5351.

doi: 10.3390/ijerph20075351

11. Thanh MT, Van Toan N, Ngoc VT, Tra NT, Giap CN, Nguyen DM. Deep learning application in dental caries detection using intraoral photos taken by smartphones. *Appl Sci*. 2022;12(11):5504.

doi: 10.3390/app12115504

12. Rizzoli A. *6 Innovative Artificial Intelligence Applications in Dentistry*; 2021. Available from: https://www.v7labs.com/blog/ai-in-dentistry [Last accessed on 2021 Oct 26].

13. Nirzu. *Dental Disease Detection from Panoramic Dental X-ray*. Available from: https://github.com/nirzu97/project-dental-disease-detection [Last accessed on 2021 Oct 26].

14. Nakai K, Wei L. Recent advances in the prediction of subcellular localization of proteins and related topics. *Front Bioinform*. 2022;2:910531.

doi: 10.3389/fbinf.2022.910531

15. Acharya A. *Guide to Image Segmentation in Computer Vision: Best Practices*; 2022. Available from: https://encord.com/blog/image-segmentation-for-computer-vision-best-practice-guide [Last accessed on 2022 Nov 07].

16. Brownlee J. *A Gentle Introduction to Object Recognition with Deep Learning*. Vol. 5. Machine Learning Mastery; 2019. p. 10. Available from: https://machinelearningmastery.com/object-recognition-with-deep-learning [Last accessed on 2022 Nov 07].

17. Fernandes AF, Dórea JR, Rosa GJ. Image analysis and computer vision applications in animal sciences: An overview. *Front Vet Sci*. 2020;7:551269.

doi: 10.3389/fvets.2020.551269

18. Chen B, Chen C, Hu J, *et al*. Computer vision and machine learning-based gait pattern recognition for flat fall prediction. *Sensors (Basel)*. 2022;22:7960.

doi: 10.3390/s22207960

19. ForouzeshFar P, Safaei AA, Ghaderi F, Hashemikamangar SS. Dental caries diagnosis from bitewing images using convolutional neural networks. *BMC Oral Health*. 2024;24(1):211.

doi: 10.1186/s12903-024-03973-9

20. Park EY, Jeong S, Kang S, Cho J, Cho JY, Kim EK. Tooth caries classification with quantitative light-induced fluorescence (QLF) images using convolutional neural network for permanent teeth *in vivo*. *BMC Oral Health*. 2023;23(1):981.

doi: 10.1186/s12903-023-03669-6

21. Prados-Privado M, Garc Villalón JC, Martínez-Martínez CH, Ivorra C, Prados-Frutos JC. Dental caries diagnosis and detection using neural networks: A systematic review. *J Clin Med*. 2020;9(11):3579.

doi: 10.3390/jcm9113579

22. Kang IA, Ngnamsie Njimbouom S, Lee KO, Kim JD. DCP: Prediction of dental caries using machine learning in personalized medicine. *Appl Sci*. 2022;12(6):3043.

doi: 10.3390/app12063043

23. Albano D, Galiano V, Basile M, *et al*. Artificial intelligence for radiographic imaging detection of caries lesions: A systematic review. *BMC Oral Health*. 2024;24(1):274.

doi: 10.1186/s12903-024-04046-7

24. Park EY, Cho H, Kang S, Jeong S, Kim EK. Caries detection with tooth surface segmentation on intraoral photographic images using deep learning. *BMC Oral Health*. 2022;22(1):573.

doi: 10.1186/s12903-022-02589-1

25. Lian L, Zhu T, Zhu F, Zhu H. Deep learning for caries detection and classification. *Diagnostics* (*Basel*), 2021;11(9):1672.

doi: 10.3390/diagnostics11091672

26. Alharbi SS, AlRugaibah AA, Alhasson HF, Khan RU. Detection of cavities from dental panoramic x-ray images using nested u-net models. *Appl Sci*. 2023;13(23):12771.

doi: 10.3390/app132312771

27. Sikri A, Sikri J, Piplani V, Thakur Y. Applications of artificial intelligence in dentistry: A narrative review. *South Asian Res J Oral Dent Sci*. 2024;6(1):1-10.

doi: 10.36346/sarjods.2024.v06i01.001

28. Zhou X, Yu G, Yin Q, Liu Y, Zhang Z, Sun J. Context aware convolutional neural network for children caries diagnosis on dental panoramic radiographs. *Comput Math Methods Med*. 2022;2022:6029245.

doi: 10.1155/2022/6029245

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Integrated sources model: A new space-learning model for heterogeneous multi-view data reduction, visualization, and clustering

Paul Fogel[1]*, Christophe Geissler[1], Franck Augé[2], Galina Boldina[3], and George Luta[4]

[1]Data Services, Mazars, Courbevoie, France

[2]Translational Precision Medicine, Sanofi, Vitry-sur-Seine, France

[3]Precision Medicine and Computational Biology, Sanofi, Vitry-sur-Seine, France

[4]Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University Medical Center, Washington, D.C., United States of America

Abstract

In machine learning, multi-view data involve multiple distinct sets of attributes ("views") for a common set of observations; when each view has the same attributes considered in different contexts, the data are said to contain multiple views of homogeneous format, which can be conceptualized as a tensor. In this article, we describe a novel approach for integrating multiple views of heterogeneous format into a common latent space using a workflow that involves non-negative matrix and tensor factorization (NMF/NTF). This approach, which we refer to as the integrated sources model (ISM), consists of two main steps: Embedding and analysis. In the embedding step, the views are transformed into matrices with common non-negative components. In the analysis step, the transformed views are combined into a tensor and decomposed using NTF. We also present a variant of ISM; the integrated latent sources model (ILSM), which offers significant advantages over ISM in terms of computational power and in cases where the views are highly unbalanced with regard to the number of attributes per view. Noteworthy, ISM can be extended to process multi-omic and multi-view datasets even in the presence of missing views. We provide a proof-of-concept analysis using five examples, including the UCI Digits (the University of California Irvine Pen-Based Recognition of Handwritten Digits) dataset, a public cell-type gene signatures dataset, and a multi-omic single-cell dataset. These examples demonstrate that, in most cases, multi-view clustering is better achieved with ISM or its variant ILSM than with other latent space approaches. We also show how the non-negativity and sparsity of the ISM model components enable straightforward interpretations, in contrast to other approaches that involve latent factors of mixed signs. Finally, we present potential applications to single-cell multi-omics and spatial mapping, including spatial imaging, spatial transcriptomics, and computational biology, which are currently under evaluation. ISM relies on state-of-the-art algorithms invoked through a simple workflow implemented in Python.

Keywords: Principal component analysis; Non-negative matrix factorization; Non-negative tensor factorization; Multi-view clustering; Canonical correlation analysis; Common principal components; Multidimensional scaling

***Corresponding author:**
Paul Fogel
(paul.fogel@mazars.fr)

Citation: Fogel P, Geissler C, Augé F, Boldina G, Luta G. Integrated sources model: A new space-learning model for heterogeneous multi-view data reduction, visualization, and clustering. *Artif Intell Health*. 2024;1(3):89-113.
doi: 10.36922/aih.3427

Received: April 16, 2024

Accepted: June 5, 2024

Published Online: July 24, 2024

1. Introduction

In machine learning, multi-view data involve multiple distinct sets of attributes ("views") for a common set of observations. In the special case where each view has the same attributes but is considered in different contexts, the data are a multidimensional array of order three that can be conceptualized as a tensor. For example, an RGB image has three color channels: Red, green, and blue, each being a non-negative two-dimensional (2D) matrix in which the intensity of the respective color is stored for each pixel. Non-negative tensor factorization (NTF) is a powerful latent space representation technique designed to analyze non-negative multidimensional arrays of order three or more. In the RGB image example, NTF captures both color and spatial information using non-negative factors, which can be used for various tasks such as image compression, enhancement, segmentation, classification, and fusion.[1]

Unfortunately, NTF cannot be applied to multi-view data when the views have heterogeneous content with distinct sets of attributes. For example, a text document can be mapped to different views, such as bag-of-words, topic modeling, or sentiment analysis, each with a different set of attributes. Another example is the transformed the University of California Irvine Pen-Based Recognition of Handwritten Digits (UCI Digits) dataset analyzed in this article. In this dataset, the original bitmaps of handwritten digits, extracted from a preprinted form, have been subjected to various transformations (e.g., Fourier, profile correlations, Karhunen-Love coefficients, pixel averages of images from 2 × 3 windows, Zernike moments, and morphological features), resulting in views with very different formats unsuitable for the direct application of NTF. Numerous algorithms have been proposed for handling such heterogenous multi-view data, some of which have become popular in the machine learning community. For example, the MVLEARN package uses the scikit-learn API to make it easily accessible to Python users,[2] while the Multi-Omics Factor Analysis (MOFA and MOFA+) Bioconductor packages[3,4] are widely used for the analysis of multi-omics datasets. However, since these algorithms assume a heterogeneous data structure, they do not incorporate NTF's explicit factorization of a three-dimensional (3D) array.

Other methods first convert each view into a similarity matrix between the observations, using techniques such as cosine similarity, Euclidean distance, transition probability, or self-representation learning. Since all views refer to the same observations, the similarity matrices have the same shape regardless of the view they originate from, resulting in a tensor of similarity matrices. Multi-view clustering (MVC) is performed on these similarity matrices, sometimes using tensor-based approaches.[5,6] However, these clustering approaches cannot be applied to other tasks, such as dimensionality reduction. This is because the representations of such similarity matrices do not project the data from multiple views into a common latent space with a small number of common attributes, such as underlying factors or concepts.

Another strategy, which allows the use of tensor decomposition techniques, starts by selecting representative points from the data, known as anchor points. These anchor points act as intermediaries to derive transition probabilities from samples to clusters. Within each view, an anchor graph estimates the probability transition matrix from the observations to the anchor points, typically by imposing a sum-to-one constraint on non-negative similarity indices over all anchor points for each point. Within each view, the probability transition matrices from anchor points to clusters and from observations to clusters need to be estimated, together with the clustering labels of the observations. For this purpose, NTF is applied with an orthogonality constraint on the cluster indicator matrices. A shadow p-norm constraint ensures that the cluster labels are consistent across views.[7-9] This approach is primarily designed for MVC, as it requires a special algorithm to select the anchor points that are best distributed across the clusters. It should be noted that many MVC approaches do not involve tensor decomposition techniques. For example, fuzzy-model-based robust clustering on multivariate *t*-mixture distributions (F-MB-T)[10] uses a *t*-mixture model in the expectation-maximization algorithm, resulting in more robust clustering. Unsupervised multi-view K-means or fuzzy C-means[11,12] consider a K-means-like membership architecture across different views. To eliminate the need for a predefined number of clusters, these methods add penalty terms to construct an unsupervised regularization structure. Starting with each data point forming its own cluster, an agglomerative process allows such approaches to be initialization-free.

This article introduces the integrated sources model (ISM), which allows NTF to analyze non-negative heterogeneous views, albeit indirectly, by means of a preliminary embedding of the data in a latent space common to all views. To this end, each view is subjected to non-negative matrix factorization (NMF), using a simple process that ensures consistency between the NMF components across all views. This consistency ensures that the embedded views share the same (synthetic) attributes, forming a non-negative 3D array that can be analyzed by NTF. Our goal in pursuing this strategy is to directly benefit from the proven performance and convergence properties of the NMF and NTF algorithms,

whose availability in powerful MATLAB, Python, or R packages ensures scalability, as will be shown in the results section (Section 3), and accessibility for the vast majority of the machine learning community. In addition to the NTF components, a view-mapping matrix is estimated to obtain an interpretable link between the dimensions of the latent space and the original attributes from each view. It is worth noting that there are some commonalities between ISM and the anchor-based approaches mentioned above, which are discussed further in the discussion section (Section 4).

The ISM belongs to the class of multi-view latent space representation methods,[13-23] which aim to capture underlying factors or concepts that characterize the data in the latent space while filtering out noise and redundancy. For MVC applications, performing cluster analysis in the latent space generally results in more accurate and consistent cluster partitioning.[24] It is noteworthy that these approaches allow newly collected data (i.e., data that are not part of the data used to train/learn the model) to be embedded in the latent space, thus extending beyond the purpose of MVC. Some of the latent space representation methods generate NMF-based latent factors[21,23] using regularization parameters that ensure sparsity and consistency between model parameters across different views. The originality of ISM lies in its simple workflow involving NMF and NTF steps. As a result, ISM produces latent factors whose interpretation is greatly facilitated by the non-negativity of the attribute loadings that define them, since they cannot cancel each other out. The interpretability of latent factors is of critical importance if they are to be used by an investigator as a follow-up tool, for example, in a clinical trial comprising several surveys with heterogeneous content.

Finally, we show that embedding the views in a 3D array has broader implications in a number of areas, such as parallelization, federated computing, and distributed computing, further illustrating the scalability and versatility of ISM, which extends well beyond the scope of multi-view data analysis.

2. Data and methods

2.1. Data

Five datasets, all with labeled observations, are considered in this article. The labeling will be used for the evaluation of the clustering performance of ISM and other methods. Details of the five datasets are as follows:

(i) UCI Digits dataset: This dataset, available in the UCI machine learning repository[25] (https://archive.ics.uci.edu/dataset/72/multiple+features), contains six heterogeneous views of handwritten digits: 76 Fourier coefficients of the character shapes, 216 profile correlations, 64 Karhunen-Love coefficients, 240-pixel averages of the images from 2 × 3 windows, 47 Zernike moments, and six morphological features. Each class of digits (0 – 9) contains 200 labeled examples.

(ii) Signature 915 data: This dataset is available in the GitHub repository (https://github.com/Advestis/adilsm/tree/main/examples/data) in the file "abis_915.csv." It comprises expression data of 915 marker genes in four patients and 16 cell types). There are four views of 915 gene markers (one view per patient) measured across 16 different cell types.[26]

(iii) Reuters dataset: Available in the GitHub repository (https://github.com/mbrbic/Multi-view-LRSSC/tree/master/datasets) in the file "Reuters.mat," Reuters dataset contains features of documents in five different languages over a common set of six categories.[27] All documents are represented in the bag-of-words format. Each of the six classes contains 100 documents, resulting in a dataset of 600 documents. The word counts in each view are 21,526, 24,892, 34,121, 15,487, and 11,539 words, respectively.

(iv) Prokaryotic phyla dataset: Found in the GitHub repository (https://github.com/mbrbic/Multi-view-LRSSC/tree/master/datasets) in the file "prokaryotic.mat," prokaryotic phyla dataset contains 551 prokaryotic species described with heterogeneous multi-view data,[28] including textual data (438 features), proteome composition encoded as relative frequencies of amino acids (three features), and gene repertoire (393 features) encoded as presence/absence indicators of gene families in a genome. Each provided view contains the principal components explaining 90% of the variance.[22] Each species in the dataset is labeled with its phylum, resulting in four unbalanced categories ranging from 35 to 313 species.

(v) TEA-seq multi-omic single-cell dataset: This dataset, available in the figshare repository (https://figshare.com/s/1b13e12f33e83fff7e0e) in the file "tea_preprocessed.h5mu," consists of human peripheral blood mononuclear cells. It includes paired profiling of scRNA-seq (2,500 features), scATAC-seq (15,000 features), and surface proteins (46 features).[29] As the dataset did not come with cell annotations, an annotation was derived from the clustering of cells using MOFA+ with 15 components,[21] resulting in seven major cell types: CD4 effector and memory T cells, B cells, CD4+ naïve T cells, monocytes, CD8+ T cells, Mucosal-associated invariant T (MAIT) cells, and natural killer (NK) cells.

Of note, the UCI Digits and Signature 915 datasets cover both aspects of sparsity (because the Signature 915

dataset contains the expression of marker genes) and redundancy (because the UCI Digits dataset contains redundant information in the nature of the images). For this reason, special emphasis is placed on the analysis of these datasets.

2.2. Methods

2.2.1. Outline of ISM and comparison with other latent space approaches

Before delving into the details of the ISM workflow, we present the main underlying ideas with an illustrative figure (Figure 1A) and compare ISM with other latent space approaches (Figure 1B). The different views are represented by heatmaps on the left side of both panels, with attributes on the vertical axis and observations on the horizontal axis.

(a) ISM

In the central part of Figure 1A, each non-negative view *Xv* is decomposed into the product of two non-negative matrices, H_v and W_v, using NMF. Each W_v matrix corresponds to the transformation of a particular view *v* to a latent space common to all transformed views. ISM ensures that the transformed views, W_v, share the same number and type of latent attributes, as explained in the detailed description. This transforming process, which we call embedding, results in a 3D array, or tensor. The corresponding H_v matrices contain the loadings of the original attributes on each component. We refer to these matrices as the mapping between the original and transformed views.

In the right part of Figure 1A, the 3D array is decomposed into the tensor product of three matrices: W^*, H^*, and Q^* using NTF. W^* contains the meta-scores – the single transformation to the latent space common to all views. H^* and Q^* contain the loadings of the latent attributes and views, respectively, on each NTF component. Each row of Q^* is represented by a diagonal matrix, where the diagonal contains the loadings for a particular view. This allows for each view of the tensor to translate the tensor product into a simple matrix product $W^*(H^*Q_v^*)^T$, as seen in Figure 1A.

(b) Other latent space approaches

In the right part of Figure 1B, each view *v* is decomposed into the product of two matrices, H_v and W, using the latent space method algorithm. As with ISM, W contains the meta-scores – the single transformation in the latent space common to all views.

(c) Comparison between ISM and other latent space approaches

If we multiply each mapping matrix H_v by H^*Q^* in Figure 1A, we obtain a representation similar to that in Figure 1B. This shows that ISM belongs to the family of latent space decomposition methods. However, view loadings are a constitutive part of ISM, whereas in other models, they are derived separately. For example, the MOFA+ method uses variance decomposition by factor.[3]

(d) Important implications of ISM's preliminary embedding

As will be detailed in the workflow description, ISM begins by applying NMF to the concatenated views. Importantly, NMF can be applied to each view X_v separately, leading to view-specific decompositions $X_v = W_v^{nmf} H_v^{nmf\,T}$ before ISM itself is applied to the *m* NMF-transformed views W_v^{nmf}. In this case, the view mapping returned by ISM, H_v^{ism}, refers to the NMF components of each W_v^{nmf}. However, by embedding the W_v^{nmf} in a 3D array, ISM allows H_v^{ism} to be mapped back to the original views through simple chained matrix multiplication such that: $X_v = W^* H_v^T$ with $H_v = H_v^{nmf} H_v^{ism} H^* Q_v^*$. We refer to this alternative approach as integrated latent source model (ILSM). As shown in the results (Section 3) and discussion (Section 4) sections, ILSM offers important advantages in several respects.

2.2.2. Compared methods

In this article, we compare ISM and ILSM with multi-view multidimensional scaling (MVMDS),[2,14] MOFA+,[3,4] group factor analysis (GFA),[18] and Multi-Omics Wasserstein inteGrative anaLysIs (MOWGLI).[21] Below is a brief description of each of these methods.

(a) MVMDS: After computing and double-centering the Euclidean distance matrices for each of the views, MVMDS estimates the common principal components of the matrices in a manner similar to the generalization of principal component analysis (PCA) for multiple covariance matrices

(b) MOFA+ and GFA: Both models are formulated in a probabilistic Bayesian framework, where prior distributions are placed on all unobserved variables of the model, using a standard normal prior for the factors W and sparsity priors for the mapping matrices H_v

(c) MOWGLI: This model is a multi-view generalization of NMF, using optimal transport instead of the Frobenius cost function and regularization parameters that ensure sparsity and consistency between model parameters across different views. A sum-to-one constraint is applied to the common factors W to give them a probabilistic interpretation.

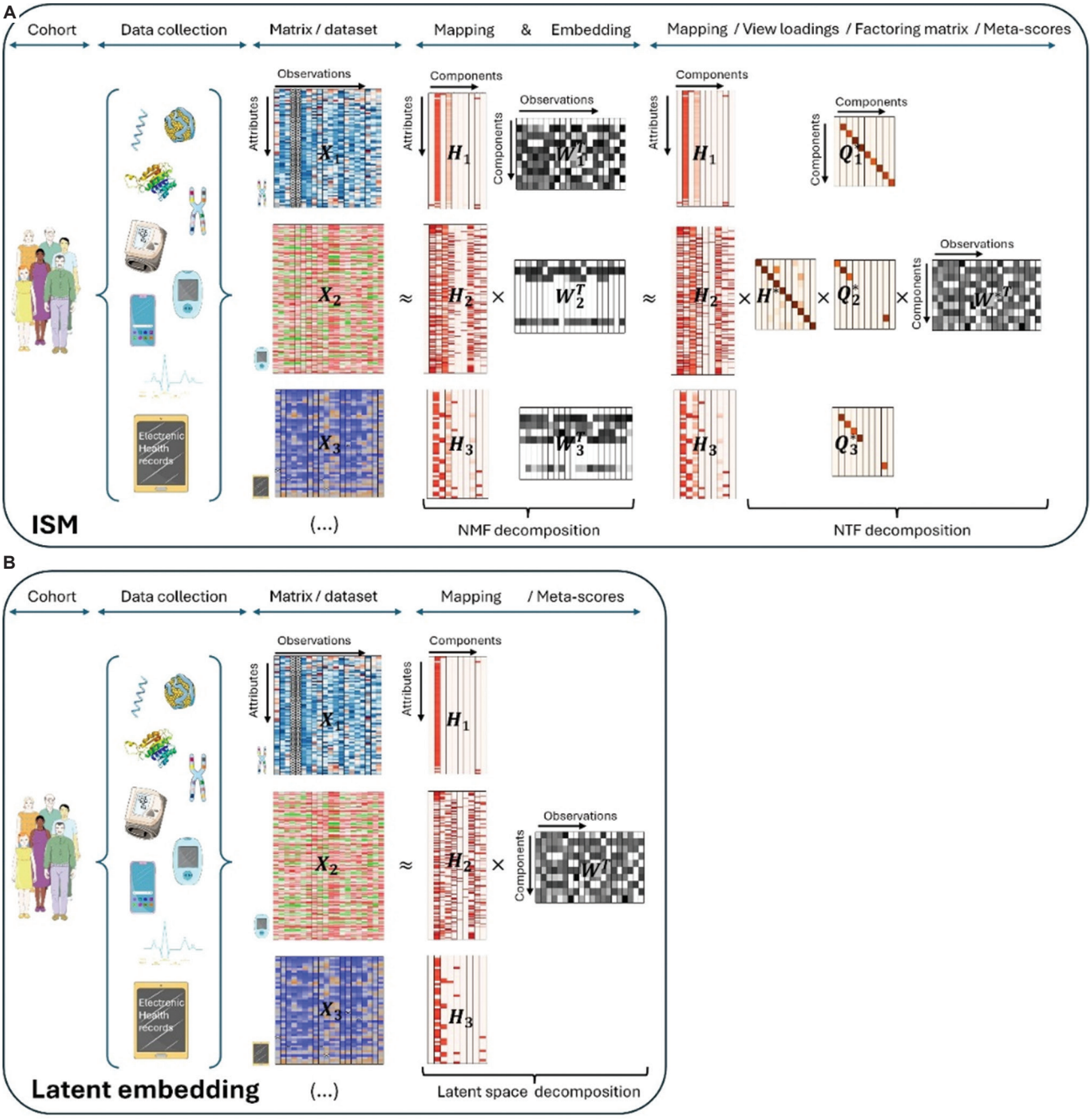


Figure 1. Comparison between integrated sources model (A) and other latent space approaches (B). Note: The image was created using elements provided by Servier Medical Art (https://smart.servier.com/citation-sharing/).
Abbreviations: NMF: Non-negative matrix factorization; NTF: Non-negative tensor factorization.

2.2.3. Detailed workflows

In this section, we present three workflows. The first workflow consists of training the ISM model to generate a latent space representation and view-mapping. The second workflow enables the projection of new observations obtained in multiple views into the latent space. The third workflow contains the detailed analysis steps for each example.

(a) Workflow 1: Latent space representation and view-mapping

The training of the ISM model can be divided into five units, as described in Figure 2. The first four process

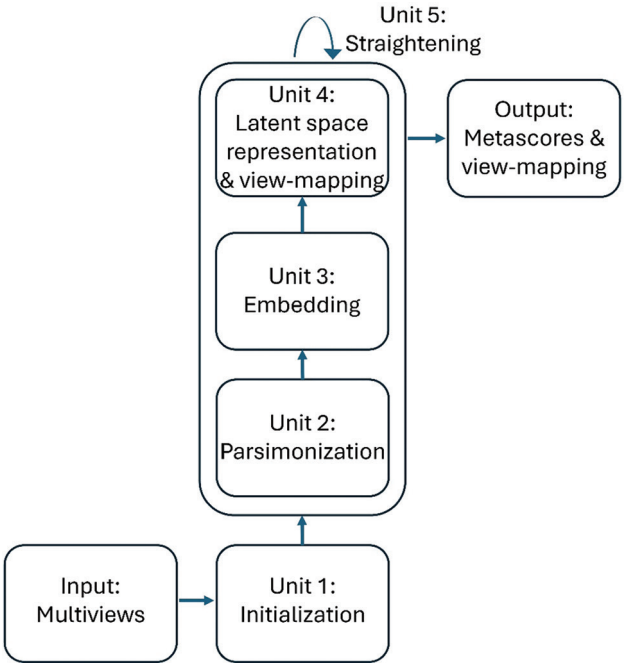


Figure 2. Training of the integrated sources model

units enable the discovery of the latent space within an "embedding" space. Once the latent space is identified, it is assimilated with the embedding space. During the fifth "straightening" unit, the latent space remains fixed, while the sequence of Units 3, 4, and 2 is repeated to further parsimonize the view-mapping until the degree of sparsity remains unchanged. The theoretical foundations of combining NMF and NTF during the embedding and latent space discovery steps are discussed at the end of this section. The sizes of the embedding space and the latent space are discussed in the section describing the third workflow.

(i) Unit 1: Initialization

An NMF is first performed on the matrix X of the m concatenated views X_v, $1 \le v \le m$, resulting in the decomposition: $X = WH^T + E, W \in \mathbb{R}_+^{n \times d_e}, H \in \mathbb{R}_+^{d \times d_e}$, $E \in \mathbb{R}^{n \times d}$ where W represents the transformed data, the columns of H contain the loadings of the $d = \sum_{v \le m} d_v$ attributes across all views on each component, d_e is the embedding size, and n is the total number of observations.

Unit 1. Initialization

Input: m views $\{X_1, , X_m\}$, $X_v \in \mathbb{R}_+^{n \times d_v}$ where n is the number of rows common to all views and d_v is the number of columns in the v^{th} view (it is assumed for each column that its values lie between 0 and 1 after normalization by the maximum row value).

Output: Factoring matrices $W \in \mathbb{R}_+^{n \times d_e}$, $H \in \mathbb{R}_+^{d \times d_e}$ where d_e is the embedding dimension, and is the sum of the number of columns in all views and $d = \sum_{v \le m} d_v$ the matrix of concatenated views X.

1: Concatenate the m views: $X = \left[X_1, \cdots, X_m\right], X \in \mathbb{R}_+^{n \times d}$;

2: Factorize X using NMF with d_e components:

$X = WH^T + E, W \in \mathbb{R}_+^{n \times d_e}, H \in \mathbb{R}_+^{d \times d_e}, E \in \mathbb{R}^{n \times d}$;

(ii) Unit 2: Parsimonization

The initial degree of sparsity of H is crucial to prevent the embedding dimensions from being overly distorted between the different views during the embedding process, as will be seen in the next section. This is achieved by applying a hard threshold to each column of the H matrix. The threshold is based on the reciprocal of the Herfindahl-Hirschman index (HHI),[30] which provides an estimate of the number of non-negligible values in a non-negative vector.

For columns with strongly positively skewed values, the use of the L2 norm for the estimate's denominator can lead to excessively sparse factors, which in turn can lead to an overly large approximation error during embedding. Therefore, the estimate is multiplied by a coefficient whose default value was set at 0.8, after testing with simulated datasets using the simulation framework described in Fogel *et al.*[31]

Unit 2. Parsimonization

Input: Factoring matrix

Output: Parsimonized factoring matrix $H \in \mathbb{R}_+^{d \times d_e}$ (since the initial H is not used outside parsimonization, we use the same symbol for the sake of simplicity).

1: for each component h_k of H do

2: Calculate the reciprocal of the Herfindahl-Hirschman Index to estimate the number of non-negligible entries in h_k:

$$\tau_k = \frac{\left(\sum_{i \le d} h\left[i,k\right]\right)^2}{\sum_{i \le d} h\left[i,k\right]^2} = \frac{h_{k1}^2}{h_{k2}^2};$$

3: Enforce sparsity on hk using hard thresholding:
If rank $(h\,[i,k]) < \tau_k \times \lambda$ then set $h\,[i,k] = 0$ where λ is a sparsity parameter ($0 < \lambda < 1$, the default value $\lambda = 0.8$ was chosen as it led in many trials to better results than the original index τ_k, which may be a too strict filter);

4: end for

(iii) Unit 3: Embedding

The matrices W and H are further updated along each view, yielding matrices $W_v \in \mathbb{R}_+^{n \times d_e}$ of common shape (number of observations n × factorization rank d_e) corresponding to the transformed views.

NMF multiplicative updates are used during view matching to leave the zeros in the primary H matrix unchanged. Further optimizations of the simplicial cones $H_v \in \mathbb{R}_+^{d_v \times d_e}$ for each view v are therefore limited to the non-zero loadings so that they remain tightly connected. This ensures that the transformed views W_v, $v \leq m$, form a tensor. Multiplicative updates usually start with a linear rate of convergence, which becomes sublinear after a few hundred iterations.[32] By default, the number of iterations is set to 200 to ensure a reasonable approximation to each view, as required for the latent space representation described in the next section.

Unit 3. Embedding

Input: m views $\{X_1, \ldots, X_m\}$ and factoring matrices $W \in \mathbb{R}_+^{n \times d_e}$, $H \in \mathbb{R}_+^{d \times d_e}$.

Output: view-specific factoring matrices $W_v \in \mathbb{R}_+^{n \times d_e}$, $H_v \in \mathbb{R}_+^{d_v \times d_e}$ and tensor $\boldsymbol{\mathcal{A}}$.

1: for each view v do

2: Define $H_v \in \mathbb{R}_+^{d_v \times d_e}$ as the part of H corresponding to view v;

3: Factorize X_v into view-specific $W_v \in \mathbb{R}_+^{n \times d_e}$ and using $H_v \in \mathbb{R}_+^{d_v \times d_e}$ NMF multiplicative updating rules and initialization matrices $W \in \mathbb{R}_+^{n \times d_e}, H_v \in \mathbb{R}_+^{d_v \times d_e}$:

$X_v = W_v H_v^T + E_v, W_v \in \mathbb{R}_+^{n \times d_e}, H_v \in \mathbb{R}_+^{d_v \times d_e}, E_v \in \mathbb{R}^{n \times d_v}$;

4: Normalize each component of W_v by its maximum value and update H_v accordingly;

5: Define tensor slice: $\boldsymbol{\mathcal{A}}(:,:,v) = W_v$;

6: end for

(iv) Unit 4: Latent space representation and view-mapping

The resulting tensor $\boldsymbol{\mathcal{A}}$ is analyzed using NTF, which leads to the decomposition: $\boldsymbol{\mathcal{A}} = W^* \otimes H^* \otimes Q^* + \varepsilon$ where $W^* \in \mathbb{R}_+^{n \times d_l}$, $H^* \in \mathbb{R}_+^{d_e \times d_l}$, $Q^* \in \mathbb{R}_+^{m \times d_l}$, $\varepsilon \in \mathbb{R}_+^{n \times d_e \times m}$, and d_l is the dimension of the latent space. The components W^*, H^*, and Q^* enable the reconstruction of the horizontal, lateral, and frontal slices of the embedding tensor. The loadings of the views on each component are contained in the matrix Q^*. The integrated multiple views, or meta-scores, are contained in the matrix W^*. The matrix H^* represents the latent space in the form of a simplicial cone contained in the embedding space. Finally, the view-mapping matrix H is updated by applying steps 3 – 8 of Unit 4. Its sparsity is ensured by further applying Unit 2 (parsimonization).

Unit 4. Latent space representation and view-mapping

Input: view-specific factoring matrices $H_v \in \mathbb{R}_+^{d_v \times d_e}$ and tensor $\boldsymbol{\mathcal{A}}$.

Output: NTF factors $W^* \in \mathbb{R}_+^{n \times d_l}, H^* \in \mathbb{R}_+^{d_e \times d_l}, Q^* \in \mathbb{R}_+^{m \times d_l}$ and view-mapping matrix $H \in \mathbb{R}_+^{d \times d_l}$.

1: Define view-mapping matrix $H \in \mathbb{R}_+^{d \times d_e}$ as the concatenation of $H_v \in \mathbb{R}_+^{d_v \times d_e}$;

2: Factorize $\boldsymbol{\mathcal{A}}$ using NTF with d_l components: $\boldsymbol{\mathcal{A}} = W^* \otimes H^* \otimes Q^* + \varepsilon$ where $W^* \in \mathbb{R}_+^{n \times d_l}, H^* \in \mathbb{R}_+^{d_e \times d_l}, Q^* \in \mathbb{R}_+^{m \times d_l}, \mathrm{E} \in \mathbb{R}_+^{n \times d_e \times m}$;

3: Update view-mapping matrix $H \in \mathbb{R}_+^{d \times d_l}$: $H \leftarrow HH^*$;

4: for each view v do

5: Update H_v: $\mathrm{H}_v \leftarrow \mathrm{H}_v \circ \mathrm{Q}^*$ [v, :];

6 end for

7: Update view-mapping matrix $H \in \mathbb{R}_+^{d \times d_l}$ as the concatenation of updated H_v;

8: Parsimonize view-mapping matrix $H \in \mathbb{R}_+^{d \times d_l}$ by applying Unit 2;

(v) Unit 5: Straightening

The sparsity of the view-mapping matrix H can be further optimized together with the meta-scores W^* and the view-loadings Q^* by repeating Units 3, 4, and 2 until the number of zero entries in H remains unchanged. To achieve this, the embedding is restricted to the latent space defined by the simplicial cone formed by H^*. In this simplified embedding space, H^* becomes the identity matrix I_{d_l} when the updating process of W^*, H^*, and Q^* starts. In other words, the embedding and latent spaces are assimilated during the straightening process. Optionally, for faster convergence, H^* can be fixed to I_{d_l}, at the cost of a slightly higher approximation error, as observed in simulated experiments, due to only small deviations from I_{d_l}.

Unit 5. Straightening

Input: X, $\boldsymbol{\mathcal{A}}$, H, W^*, H^*, Q^*.

Output: NTF factors W^*, H^*, Q^* and updated view-mapping matrix H.

1: $H^* = I_{d_l}$ Set where d_l is the size of the latent space;

2: do until the number of 0-entries in H remains unchanged

3: Apply Unit 3 to embed X using the embedding size $d_e = d_l$, initialization matrices $W^* \in \mathbb{R}_+^{n \times d_l}$ and view-mapping matrix $H \in \mathbb{R}_+^{d \times d_l}$ found in the previous iteration;

4: Apply Unit 4 to factorize $\boldsymbol{\mathcal{A}}$ and update the view-mapping matrix $H \in \mathbb{R}_+^{d \times d_l}$, using embedding size $d_e = d_l$, initialization matrices W^*, H^*, Q^* obtained in the previous iteration and fixed $H^* = I_{d_l}$;

5: end for

(vi) Theoretical foundations of combining non-negative matrix and tensor factorization

From a more theoretical perspective, NMF estimates, for each view, the transformed data in the form of a matrix W_v and a view-mapping matrix H_v, which allows the reconstruction of the original view. Following a geometrical interpretation from Donoho and Stodden,[33] we consider

the simplicial cone Γ_{W_v} contained in the positive orthant of R^n and generated by the columns of W_v. In this simplicial cone, each view-attribute corresponds to a point, with coordinates found in the corresponding row of H_v. To identify a consensus simplicial cone between the Γ_{W_v}, NTF decomposes the tensor formed by the W_v into a sum of rank-1 tensors (Unit 4). However, for such a decomposition to be meaningful, the dimensions defined by the columns of the W_v must be consistent from one view to another. This implies a strong overlap between the simplicial cones Γ_{W_v}. Such consistency is achieved by the multiple zeros found across the columns of H_v when starting the embedding process (Unit 3). These are "inactive" attributes, as their zero status cannot be changed by multiplicative update rules. They can be interpreted as anchors ensuring that the W_v do not deviate significantly from their common ancestor W, estimated in the preliminary NMF over concatenated views (Unit 1). The parsimonization process (Unit 2) is designed to ensure that there will be a sufficient number of anchor attributes to rein in the multiplicative updates.

(b) Workflow 2: Projection of new observations

For new observations Y comprising k views, $k \leq m$, ISM parameters H^*, Q^*, and the view-mapping matrix H can be used to project Y onto the latent ISM components, as described in Workflow 2.

Workflow 2. Projection of new observations

Input: New observations Y (k views, $k \leq m$),

NTF factors H^*, Q^* and mapping matrix H.

Output: Estimation of Y^*.

1: Disregard any views in Q^*, H that are absent in Y;

2: Apply Unit 3 of Workflow 1 to embed Y with W initialized with ones and with fixed mapping matrix H;

3: Apply step 2 of Unit 4 of Workflow 1 to calculate W^* with fixed NTF factors H^*, Q^* and define the projection of Y on the latent space as $Y^* = W^*$;

Abbreviation: NTF: Non-negative tensor factorization.

(c) Workflow 3: Proof of concept analysis

Each dataset is analyzed using ISM, ILSM, NMF, MVMDS, GFA, MOFA+, and MOWGLI. PCA is also applied to the concatenated views of the UCI Digits and Signature 915 datasets, mainly to show the added value of alternative approaches over this widely used method.

To facilitate interpretation, the transformed data are projected onto a 2D map before being subjected to K-means clustering, where k is the known number of classes (K-means clustering was chosen for its versatility and simplicity, as it only requires the number of clusters to be found, and this number is known for our example datasets). Within each cluster, the class that contains the majority of the points, that is, the main class is identified. If two clusters share the same main class, they are merged unless they are not contiguous (the ratio of the distance between the centroids to the intra-cluster distance between points >1). In this case, the non-contiguous clusters are excluded because they are assigned to the same class, which should appear homogeneous in the representation. Similarly, any cluster that does not contain an absolute majority is not considered clearly representative of the class to which it is assigned and is excluded from the study. A global purity index is then calculated for the remaining clusters using Workflow 3. To enhance clarity, the clusters are visualized using 95% confidence ellipses, while the classes are represented using distinct colors. In addition to the proportion of classes retrieved and the global purity index, the adjusted rand index (ARI),[34] normalized mutual information (NMI) index,[35] and Fowlkes-Mallows score (FMS)[36] are also included, along with the factor specificity index (FSI) and view-mapping sparsity (VSI) defined as follows:

The FSI reflects the level of factor specificity with respect to a given class: A value close to 1 means that only one factor contributes significantly to the explanation of the class; while a value close to 0 means that the class is explained by a large number of factors. This index was proposed in Huizing *et al.*,[21] but in its original definition, it measures the level of specificity of each factor relative to the class. The FSI is defined as the ratio of the maximum specificity observed across all factors over the number of significant factors. To estimate the number of significant factors, we use the inverse HHI of all factor indices.

The VSI reflects the level of the sparsity of the mapping matrix H. To obtain the VSI: (i) Estimate, for each view and each ISM component, the number of significant loadings, using the inverse HHI; (ii) for each view, define the view-sparsity as the average sparsity over all ISM components; and (iii) define VSI as the average view-sparsity over all views.

Multidimensional scaling is applied to achieve the 2D map projection. MDS uses a simple metric objective to find a low-dimensional embedding that accurately represents the distances between points in the latent space.[37] MDS is, therefore, agnostic to the intrinsic clustering performances of the methods that we want to evaluate. Effective embedding methods, for example, uniform manifold approximation and projection (UMAP) or *t*-distributed stochastic neighbor embedding, are not as optimal for preserving the global geometric structure in the latent space.[38] For example, a resolution parameter needs to

be defined for the UMAP embedding of single-cell data, whereby a higher resolution leads to a higher number of clusters. In addition, the subtle differences between some cell types from one family can be smoothed out if the dataset contains transcriptionally distinct cell types from multiple families, as is the case with immune cells for the Signature 915 dataset.

Latent space methods require that the rank of the factorization is determined in advance. ISM benefits from the advantages of the NMF and NTF workflow components, that is, the choice of the correct rank is less critical than with other methods (we will come back to this point in the results [Section 3] and discussion [Section 4] sections). This allows, even if we expect some redundancy in the latent factors – for instance, due to the proximity of certain digits in the first dataset – to set the rank to the number of known classes.

The dimension of the ISM embedding space must also be determined during the discovery step. A natural choice is the dimension of the latent space since both spaces are merged at the end of the ISM workflow. Nevertheless, by examining the approximation error for an embedding dimension in the neighborhood of the chosen rank, it is possible to further optimize the ISM representation.

The rank for PCA, MVMDS, GFA, and MOFA+ is set by inspecting the scree plot of the variance ratio.

The analysis of the Signature 915 dataset also examines the biological relevance of the distance between clusters in each latent multi-view space. Of the five datasets analyzed in this article, only the Signature 915 dataset is a 3D array; therefore, NTF is also directly applied to this particular dataset.

Detailed analysis steps are provided in Workflow 3.

Workflow 3. Analysis steps

Input: 2D map projection of the data transformation in the latent space.

Output: Cluster purity index.

1: Perform K-means with k equal to the number of known classes;

2: For each cluster, identify the main class related to the cluster, that is, the class corresponding to the majority of observations in the cluster;

3: Merge contiguous clusters that refer to the same class or ignore them if not contiguous;

4: for each cluster do

5: p_1=proportion of the main class in relation to all elements in the cluster;
p_2=proportion of the main class in cluster c in relation to all elements of the same class;

6: If $p_2 < 0.5$ then

7: Disregard cluster as the main class does not constitute an absolute majority in relation to all elements of the same class;

8: Else

9: $p=p_1 \times p_2$=purity corrected for cluster representativity for the main class;

10: end for

11: Calculate the global purity=sum of corrected purities over all retained clusters, divided by the number of known classes;

2.3. Implementation

Scikit-learn[39] was used for K-means, ARI, NMI, MDS, and PCA. The mvlearn (https://pypi.org/project/mvlearn/) package was used for MVMDS. NMF and NTF were performed with the package adnmtf (https://pypi.org/project/adnmtf/). ISM was implemented in Python and was invoked from a Jupyter Python notebook available on the Advestis GitHub (https://github.com/Advestis). GFA was performed with the Python package gfa-python (https://github.com/mladv15/gfa-python). MOFA+ was performed with the Python package mofapy2 (https://github.com/bioFAM/mofapy2). Matplotlib (https://matplotlib.org/stable/tutorials/pyplot.html) was used to create the clustering figures. Treemaps were obtained with the Graph Builder platform from JMP® (Version 17.2.0. SAS Institute Inc., USA). The distinctipy package (https://pypi.org/project/distinctipy/) was used to generate colors that are visually distinct from one another.

3. Results

We first present a synthesis of the calculated metrics across all datasets (Table 1) and provide some general observations. We then present more detailed results for each dataset.

3.1. Synthesis of calculated metrics over all datasets

Based on the average index across all seven indices, ISM ranks first in the UCI Digits, Signature 915, and Reuters datasets, while ILSM ranks first in the prokaryotic dataset and the TEA-seq multi-omic single-cell dataset (although very close to ISM for the latter dataset, 0.80 vs. 0.79, respectively). It is easy to explain why ILSM performed much better than ISM on the prokaryotic dataset (0.52 vs. 0.37, respectively): Since ISM first performs a global factorization over concatenated views (Unit 1 of Workflow 1), it tends to ignore the smallest views when they are extremely unbalanced, as is the case in the prokaryotic dataset. However, when using ILSM, separate factorizations are applied to each view, and ISM itself is applied to transformed views of equal size. As a result, the original views with the smallest size are given equal weight. Among the criteria used, the proportion of classes retrieved, purity, and sparsity indices are the most discriminative. It is noteworthy that NMF performs as

Table 1. Metrics comparing latent-space methods on five datasets

Dataset	Method	Nr classes	Embedding (ISM) rank	Proportion of classes retrieved	Purity	ARI	NMI	FMS	Sparsity	Specificity	Overall
UCI DIGITS	MVMDS	10	10	0.70	0.41	0.49	0.61	0.54	0.62	0.21	0.51
	ISM	10	(9,10)	1.00	0.58	0.57	0.67	0.62	0.87	0.43	0.68
	ILSM	10	(10,10)	0.80	0.41	0.45	0.58	0.51	0.50	0.48	0.53
	GFA	10	10	0.90	0.45	0.48	0.61	0.54	0.32	0.15	0.49
	MOFA+	10	10	0.70	0.29	0.36	0.46	0.44	0.34	0.13	0.39
	MOWGLI	10	10	0.80	0.46	0.51	0.65	0.57	0.60	0.58	0.60
	PCA	10	10	0.40	0.19	0.44	0.57	0.51	0.73	0.38	0.46
	NMF	10	10	0.90	0.58	0.59	0.68	0.63	0.46	0.34	0.60
Signature 915	MVMDS	16	10	0.75	0.70	0.97	0.95	0.97	0.56	0.21	0.73
	ISM	16	(16,16)	0.88	0.72	0.98	0.95	0.98	0.93	0.83	0.90
	ILSM	16	(16,16)	0.75	0.62	0.93	0.91	0.94	0.93	0.74	0.83
	GFA	16	12	0.81	0.73	0.98	0.96	0.98	0.30	0.08	0.69
	MOFA+	16	13	0.81	0.76	0.94	0.93	0.95	0.56	0.19	0.73
	MOWGLI	16	16	0.63	0.44	0.87	0.89	0.89	0.89	0.82	0.77
	PCA	16	10	0.56	0.40	0.94	0.89	0.95	0.57	0.23	0.65
	NMF	16	16	0.81	0.55	0.94	0.89	0.95	0.91	0.88	0.85
	NTF	16	16	0.69	0.52	0.94	0.89	0.95	0.98	0.75	0.82
Reuters	MVMDS	6	4	0.50	0.19	0.19	0.30	0.37	0.93	0.28	0.39
	ISM	6	(6,6)	0.50	0.23	0.25	0.34	0.41	0.98	0.37	0.44
	ILSM	6	(6,6)	0.33	0.16	0.21	0.31	0.39	0.97	0.30	0.38
	GFA	6	3	0.17	0.03	0.01	0.09	0.39	0.94	0.21	0.26
	MOFA+	6	-	-	-	-	-	-	-	-	-
	MOWGLI	6	6	0.08	0.04	0.01	0.20	0.28	0.10	0.86	0.22
	NMF	6	6	0.33	0.14	0.21	0.32	0.41	0.96	0.36	0.39
Prokaryotic	MVMDS	4	4	0.50	0.22	0.18	0.23	0.50	0.68	0.67	0.43
	ISM	4	(4,4)	0.25	0.14	0.00	0.00	0.63	0.66	0.88	0.37
	ILSM	4	(4,4)	0.75	0.36	0.28	0.31	0.54	0.55	0.88	0.52
	GFA	4	6	-	-	-	-	-	-	-	-
	MOFA+	4	4	0.75	0.36	0.29	0.32	0.55	0.53	0.42	0.46
	MOWGLI	4	4	0.25	0.14	0.10	0.10	0.60	0.39	0.63	0.32
	NMF	4	4	0.25	0.14	0.00	0.00	0.63	0.47	0.88	0.34
TEA-seq	MVMDS	7	7	0.71	0.60	0.89	0.86	0.92	0.67	0.48	0.73
	ISM	7	(7,7)	0.71	0.57	0.87	0.84	0.90	0.76	0.88	0.79
	ILSM	7	(7,7)	0.86	0.72	0.88	0.85	0.91	0.75	0.67	0.80
	GFA	7	15	0.71	0.61	0.91	0.89	0.93	0.45	0.25	0.68
	MOWGLI	7	7	0.43	0.23	0.52	0.60	0.64	0.39	0.62	0.49
	NMF	7	7	0.71	0.61	0.88	0.86	0.91	0.70	0.90	0.80

Abbreviations: ARI: Adjusted rand index; GFA: Group factor analysis; FMS: Fowlkes-Mallows score; ILSM: Integrated latent sources model; ISM: Integrated sources model; MOWGLI: Multi-Omics Wasserstein inteGrative anaLysIs; MVMDS: Multi-view multidimensional scaling; NMF: Non-negative matrix factorization; NMI: Normalized mutual information index.

well as ILSM in the TEA-seq multi-omic single-cell data in terms of average performance (0.80). However, we will show in the detailed analysis of this dataset that ISM finds a superior representation in terms of biology. In addition, NMF retrieves only one class in the prokaryotic data due to the extreme imbalance in the number of features per view

and the fact that NMF runs on the concatenated views, thus tending to ignore the smallest ones. Although GFA and MOFA+ are closely related, MOFA+ fails to recover common factors in the Reuters dataset, while GFA fails in the prokaryotic dataset. MVMDS performs relatively well on all datasets, in most cases with lower factor sparsity and specificity than ISM or ILSM. MOWGLI could only be run on a fraction of the data for the Reuters and TEA-seq multi-omic single-cell data due to its extremely high computational time. The poor performance observed can, therefore, be attributed to the sampling itself.

3.2. Detailed results

3.2.1. UCI digits dataset

PCA, MVMDS, MOFA+, and MOWGLI use a 10-factorization rank, while GFA uses a 9-factorization rank. ISM uses a primary embedding of dimension 9 and a 10-factorization rank. The Karhunen-Love coefficients contain data with mixed signs, so the corresponding view is split into its positive part and the absolute value of its negative part when applying the non-negative approaches ISM, NMF, and MOWGLI. The clusterings of the digits are shown in Figure 3. ISM outperforms the other methods with 10-digit-specific clusters. It should be noted that NMF performs slightly better than ISM in terms of purity index, ARI, NMI, and FMS. However, digits 5 and 3 are mixed together, resulting in one less digit being recognized. PCA is far behind all other approaches, recognizing only four-digit classes.

Figure 4 shows how the views affect the individual ISM components using a treemap chart. For each component, each view corresponds to a rectangle within a rectangular display, where the size of the rectangle represents the loading of the view. It is noteworthy that some components are supported by only a few views, for example, component 1 (2 views) and component 8 (3 views), while others involve most views, for example, component 5 (6 views). As each component is associated with a digit, this emphasizes the specifics and complementarity of the image representations that are dependent on the respective digit. It is also interesting to note that for some components, the loadings of the views are diametrically opposed to the respective number of attributes. For example, for component 8, the view of 240-pixel averages has the lowest loading, while the view of six morphological features has the highest loading. This clearly shows that the views are evenly balanced regardless of their respective number of attributes when using ISM.

3.2.2. Signature 915 data

Before the analysis, each marker gene was normalized using the mean of the four highest expression values. PCA and MVMDS use a 10-factorization rank, GFA uses a 12-factorization rank, and MOFA+ uses a 13-factorization rank. ISM uses a primary embedding of dimension 16 and a 16-factorization rank. The clusterings of the marker genes are shown in Figure 5. ISM outperforms the other methods with 14 cell type-specific clusters and higher metrics.

Regarding the positioning of the clusters on the 2D map, MVMDS places classical monocyte (monocyte C) and non-classical and intermediate monocytes (monocyte NC+I) opposite of each other, contrary to all other approaches and, more importantly, against biological intuition. ISM and GFA methods outperform other methods on this dataset as they reveal close proximity between transcriptionally and functionally similar cell types of the major immune cell families. Indeed, three cell types from the myeloid lineage, including monocytes C, monocytes NC+I, and myeloid dendritic cells (mDC), are grouped together. A similar trend is observed for three cell types from the B cell family, where only ISM and GFA revealed close proximity of naïve B cells, memory B cells, and plasmablasts, out of the eight methods considered. The most challenging cell types were in the T cell family, where only ISM was able to identify clusters for three cell types (CD4+ effectors, naïve T cells, and Vδ2+ T cells [VD+] gamma delta non-conventional T cells) and place them in close proximity. VD+ gamma delta non-conventional T cells share some similarities with NK cells in terms of the expression of certain receptors, and only the ISM method was able to recognize both cell types and place them in close proximity, highlighting their similarity. The ISM method also captured subtle similarities between two types of dendritic cells, mDC, and plasmacytoid dendritic cells (pDC), which correspond to antigen-presenting cells.

Figure 6 shows the impact of the four patients on the individual ISM components using a treemap chart. In this chart, each patient corresponds to a rectangle within a rectangular display, where the size of the rectangle represents the loading of the patient. In contrast to the UCI Digits data, most components are supported by three patients (three components) or four patients (11 components). Two components involve only two patients.

The loadings of the view-mapping matrix are shown in Figure 7 using a treemap chart. Recall that each attribute of this dataset is a combination of a patient and a cell type, in which the expressions of 915 marker genes were measured. For each component, such a combination corresponds to a rectangle within a rectangular display, where the size of the rectangle represents the loading of the combination. ISM components 1 and 2 are both associated with the same cell type, pDC, while component 15 is simultaneously associated with CD8-activated, VD2-, and VD2+ cells. In the final clustering, the cluster comprising these three

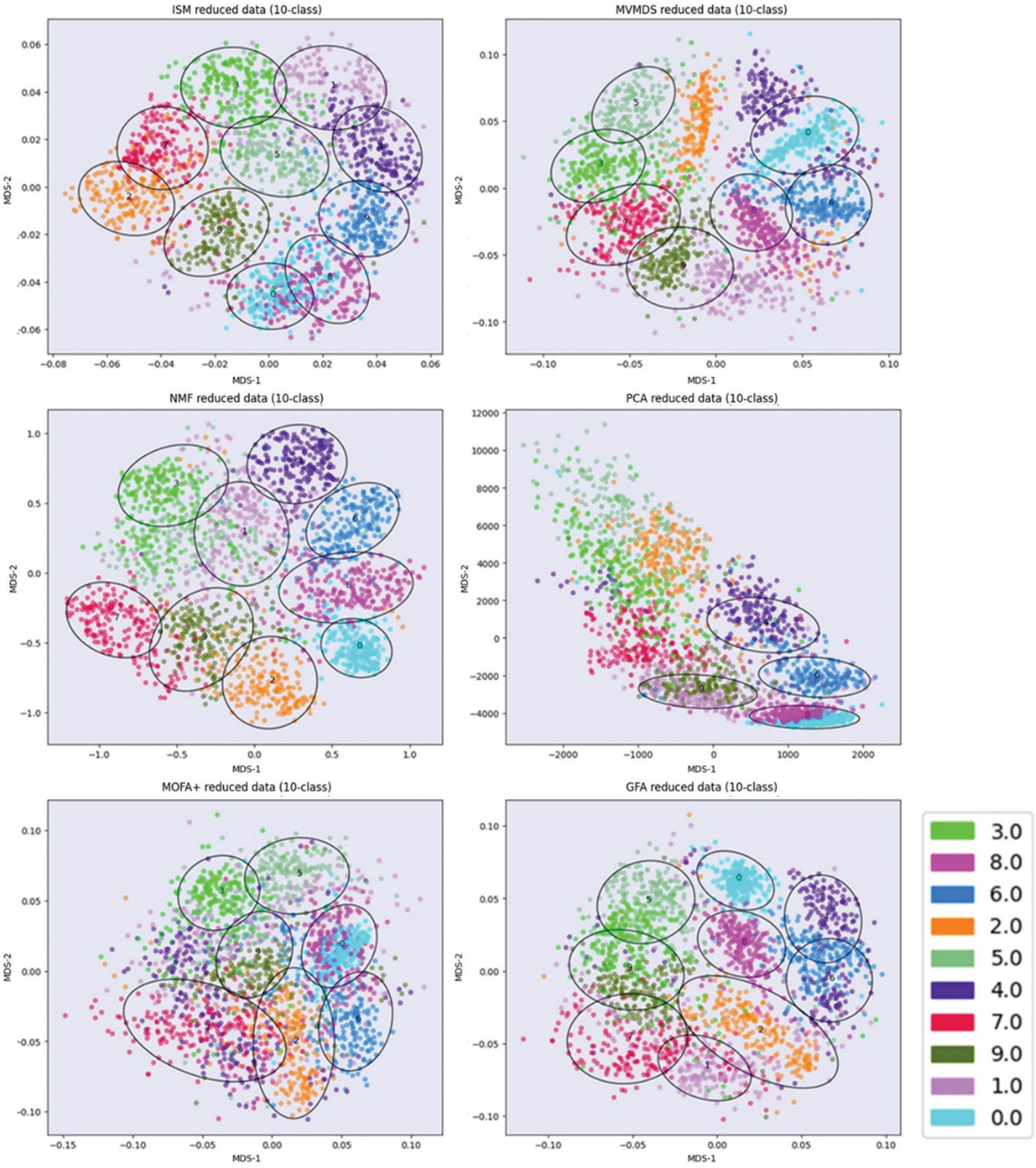


Figure 3. UCI Digits data: Clustering of digit images along ISM, MVMDS, NMF, PCA, MOFA+, and GFA components in 2D scatterplots of the MDS projection of transformed data.
Abbreviations: ISM: Integrated sources model; MDS: Multidimensional scaling; MOFA+: Multi-Omics factor analysis; MVMDS: Multi-view multidimensional scaling; NMF: Non-negative matrix factorization; NTF: Non-negative tensor factorization; PCA: Principal component analysis.

cell types has no main type and is therefore discarded, resulting in 14 identified cell types. All other components are associated with only one cell type, illustrating the sparsity and interpretability of ISM components. Notably,

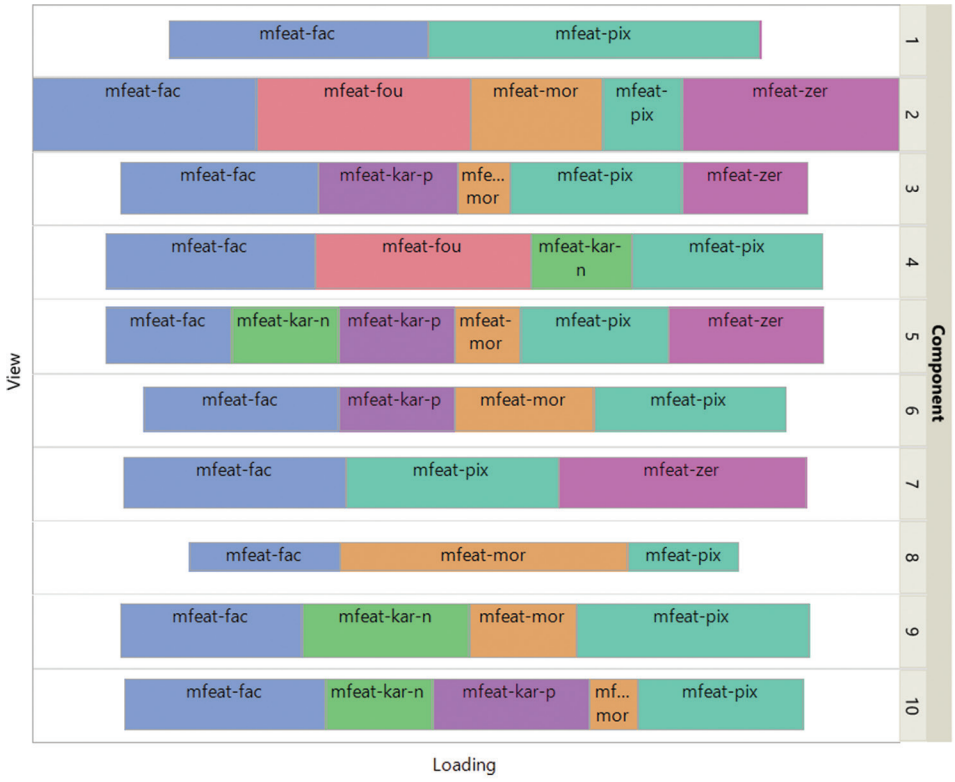


Figure 4. UCI Digits data: Treemap of integrated sources model view weights

all non-negative approaches yield a high view-mapping sparsity index (i.e., 0.93, 0.89, 0.91, and 0.98 for ISM, MOWGLI, NMF, and NTF, respectively), as opposed to mixed-sign approaches (i.e., 0.56, 0.30, 0.56, and 0.57 for MVMDS, GFA, MOFA+, and PCA, respectively).

3.2.3. Reuters data

MVMDS and MOWGLI use a 6-factorization rank, while MOFA+ and GFA use a 10-factorization rank. ISM uses a primary embedding of dimension 6 and a 6-factorization rank (equal to the number of known categories). Overall, ISM outperforms the other methods, identifying three out of six categories and achieving higher metrics, followed by MVMDS. However, all performance indices are relatively low, as previously observed in Brbic and Kopriva.[22] MOFA+ fails to identify a common structure between the different views. It should be noted that MOWGLI was performed on only 20% of the samples due to its extremely high computational time, despite using an activated graphics processing unit (GPU). The poor performance observed can, therefore, be attributed to the sampling itself.

3.2.4. Prokaryotic data

MVMDS, MOFA+, and MOWGLI use a 4-factorization rank, while GFA uses a 6-factorization rank. ILSM uses a primary embedding of dimension 4 and a 4-factorization rank (equal to the number of known categories). Since the provided views contain the principal components explaining 90% of the variance, they need to be split into their positive part and the absolute value of their negative part when applying the non-negative approaches ISM, NMF, and MOWGLI. Overall, ILSM outperforms the other methods, identifying three out of four categories (missing the category which the smallest size) and achieving higher metrics.

3.2.5. TEA-seq multi-omic single-cell data

MVMDS and MOWGLI use a 7-factorization rank, while GFA uses a 15-factorization rank. ISM uses a primary embedding of dimension 7 and a 7-factorization rank (equal to the number of categories). MOFA+ metrics are not presented for this particular dataset since the corresponding clustering was used to annotate the cells.[21] We used a UMAP projection because the size of the dataset makes MDS impractical (Figure 8).

The ILSM outperforms the other methods, identifying six out of seven cell types, missing only MAIT T cells, which are too close to CD4 effector and memory T cells to be identified as a separate cluster.

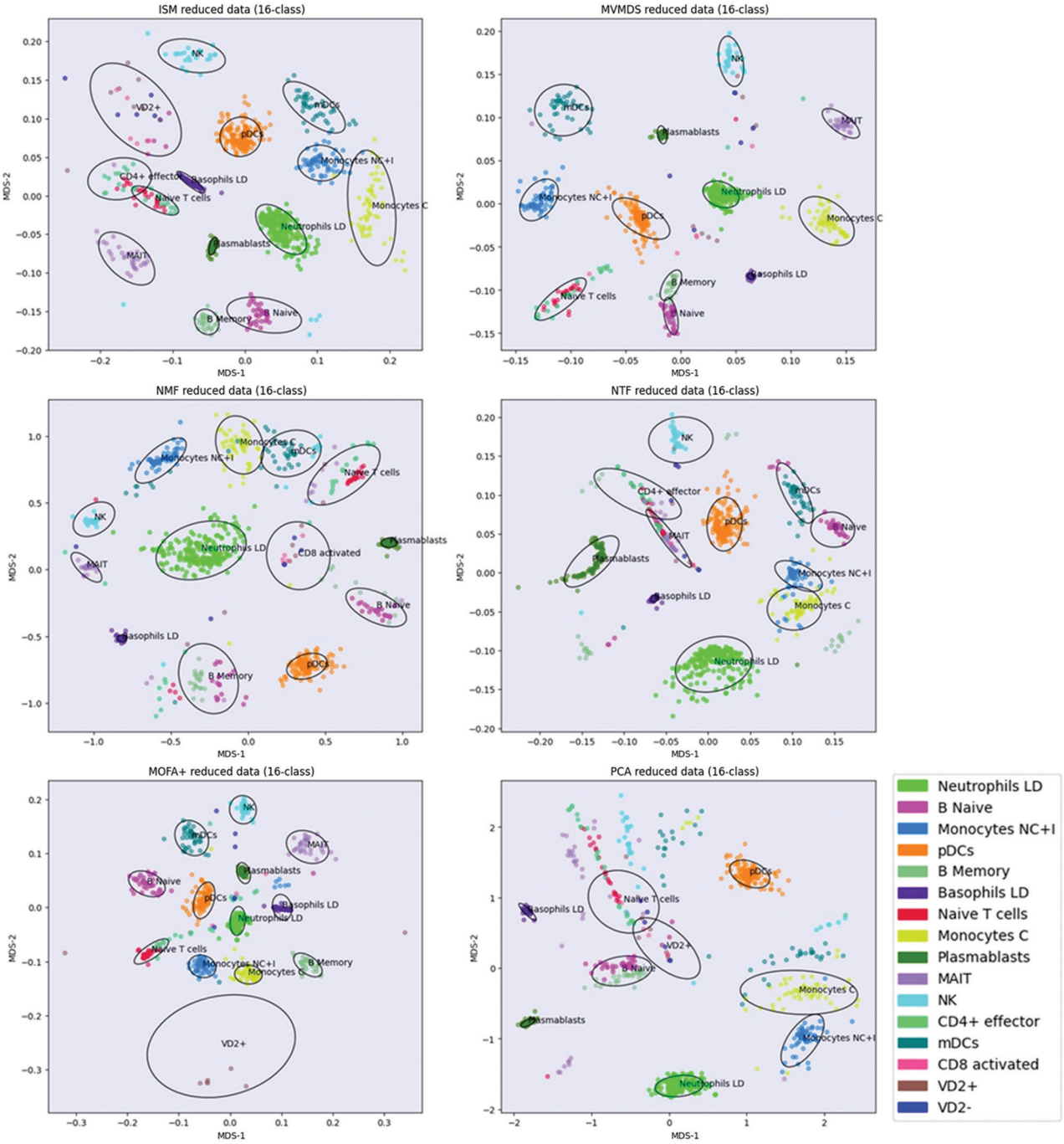


Figure 5. Signature 915 Data: Clustering of cell type marker genes along ISM, MVMDS, NMF, NTF, MOFA+, and PCA components in the 2D scatterplots of the MDS projection of the transformed data.
Abbreviations: Basophil LD: Low-density basophil; ISM: Integrated sources model; MAIT: Mucosal-associated invariant T cell; mDCs: Myeloid dendritic cells; MOFA+: Multi-Omics Factor Analysis; Monocyte C: Classical monocyte; Monocyte NC+I: Non-classical + intermediate monocytes; MVMDS: Multi-view multidimensional scaling; Neutrophil LD: Low-density neutrophil; NK: Natural killer cell; NMF: Non-negative matrix factorization; NTF: Non-negative tensor factorization; PCA: Principal component analysis; pDCs: Plasmacytoid dendritic cells; VD2: Vδ2+ T cells.

Other methods, including ISM, revealed heterogeneity in the CD4+ naive T cell population, which appeared to be split in the corresponding UMAP projections. In the ISM UMAP projection, we annotated the smaller split near

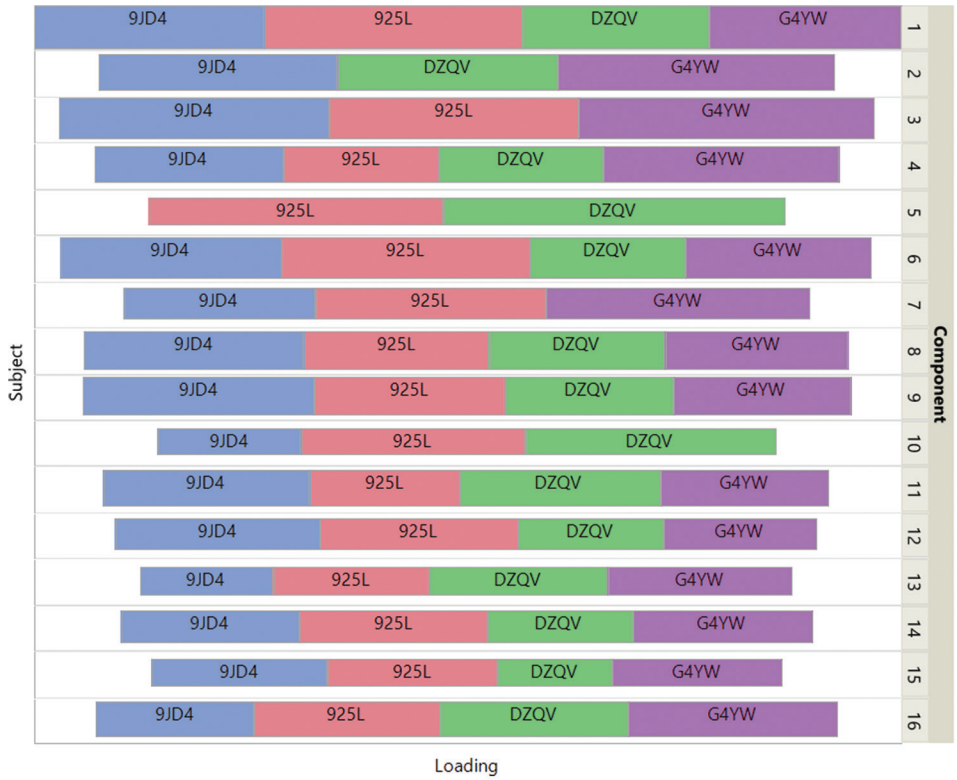


Figure 6. Signature 915 data: Treemap of integrated sources model view weights

CD8+ T cells and found that it actually represents CD8+ naive T cells. We then verified that the split observed in the other UMAP projections was also consistent with this cell type. In particular, the distance between CD8+ naive T cells and CD8+ T cells is minimal with ISM, consistent with biology. In contrast, CD8+ naive T cells are closer to CD4+ naive T cells and not to CD8+ T cells in the NMF UMAP projection, contrary to biological intuition. One possible reason is that NMF does not take advantage of the complementarity of different views by indiscriminately concatenating them.

Interestingly, and in contrast to other multi-view approaches, ISM allows the direct identification of factors and views that are discriminative with respect to a particular cell type (e.g., CD8+ naive T cells). From the factor specificities, we found two specific ISM latent factors with positive factor specificities with respect to CD8+ naive T cells (0.50 and 55, respectively). In all three multi-omic modalities, these factors have close loadings in the view-weights matrix Q^* (13.09/9.00 in the RNA-seq view, 10.82/8.04 in the ATAC-seq view, and 11.45/8.94 in the ADT view, respectively), highlighting the contribution of the three modalities to specifically distinguish CD8+ and CD4+ cell subpopulations among naive T cells.

It should be noted that MOWGLI was performed on only 20% of the samples and 20% of the scATAC-seq features due to its extremely high computational time, despite using an activated GPU. The poor performance observed can, therefore, be attributed to the sampling itself.

3.3. Further insights regarding the model

3.3.1. Model's potential dependency on embedding dimension and rank

In this section, we evaluate how ISM performance might be affected by changing the embedding dimension and the rank in the neighborhood of the chosen values. First, we examine the relative approximation error for an embedding dimension in the neighborhood of the chosen rank to select an optimal value, as described in the analysis workflow. Second, we examine the relative error, number of found classes, and purity for a rank in the neighborhood of the chosen embedding dimension.

For the UCI Digits data, where the chosen ISM rank is 10, an embedding dimension of 9 clearly minimizes the relative error—0.52 versus 0.72 or higher for other dimensions. The number of classes found and the purity index are also significantly higher (Table 2, upper part). The relative

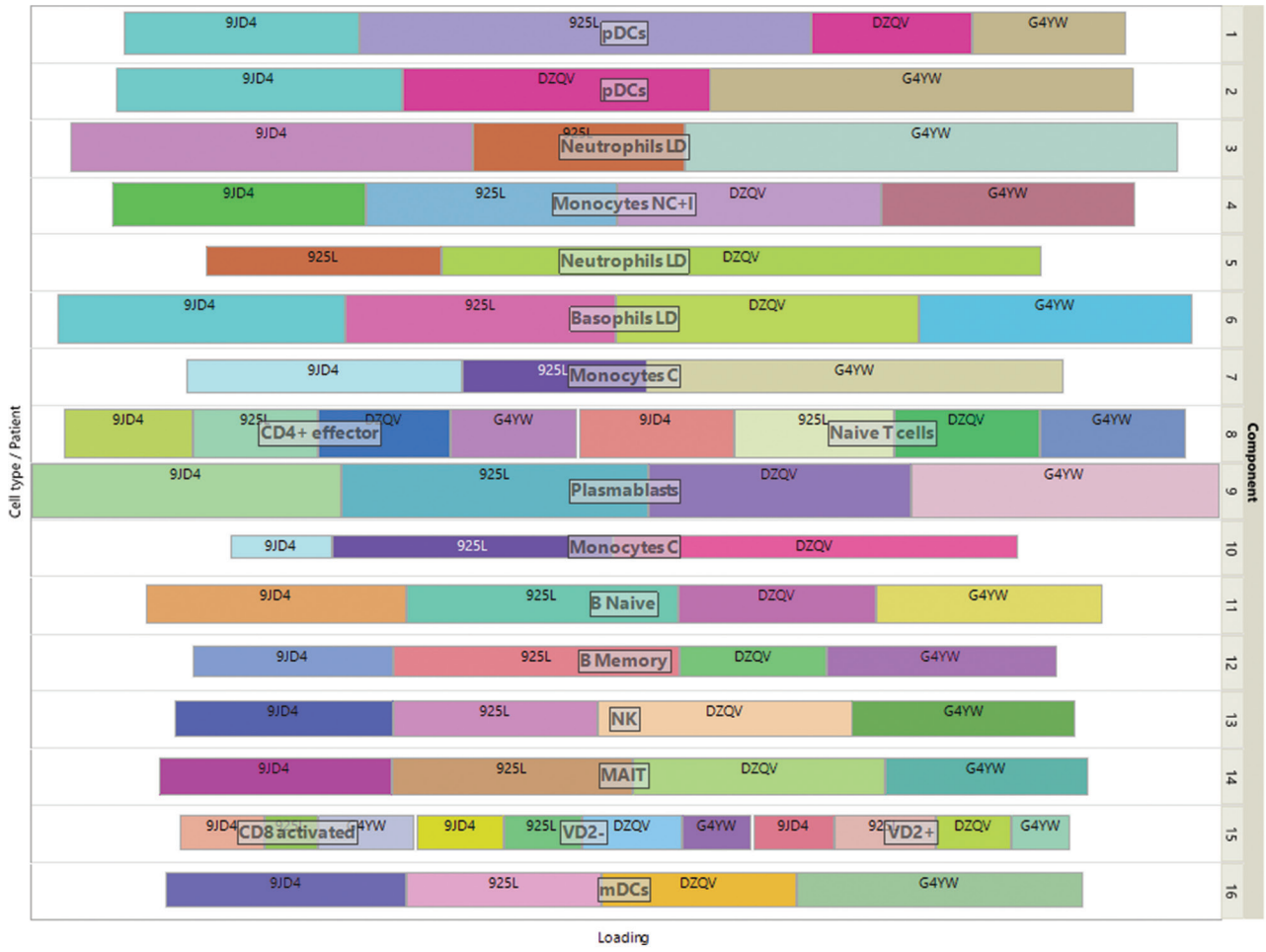


Figure 7. Signature 915 data: treemap of integrated sources model loadings of the view-mapping matrix

error associated with a rank is not as critical if it exceeds the number of known classes. Compared to a 10-rank ISM model, a 12-rank model also finds 10 classes and gives a slightly higher purity index (6.24 vs. 5.81), despite a larger relative error (0.60 vs. 0.52) (Table 2, bottom part). The final part of this section discusses this point further.

For the Signature 915 dataset, where the chosen ISM rank is 16, the relative error does not change significantly for neighboring embedding dimensions: 0.33 for a 15-embedding and 0.34 for a 17-embedding (Table 3, upper part). Choosing an embedding dimension equal to the rank is more consistent with the ISM workflow, where the embedding and latent spaces are united during the straightening process. Therefore, we chose an embedding dimension of 16. In terms of purity, a 17-rank ISM model gives results that are slightly superior to the 16-rank ISM model (Table 3, bottom part).

Overall, these results confirm that ISM provides relatively stable estimates in the neighborhood of the chosen rank, in line with its parent methods, NMF and NTF.

3.3.2. About changing the sparsity coefficient

We have already mentioned that the initial degree of sparsity of H returned by NMF is a critical part of ISM, as zero-loading attributes are anchors that maintain consistency between view components during the embedding process. However, it is extremely difficult to predict how sparse an NMF representation will be, as this depends on the dataset under analysis.[33] To ensure that a sufficient number of anchors will guide the embedding, only significant loadings are retained, while other loadings are set to 0. ISM uses the inverse of the HHI to identify significant loadings, but an additional sparsity parameter is provided to allow this index to be relaxed. This parameter is set to 0.8 by default. In this section, we examine the effect of changing this parameter in the UCI Digits and Signature 915 experiments (Tables 4 and 5, respectively).

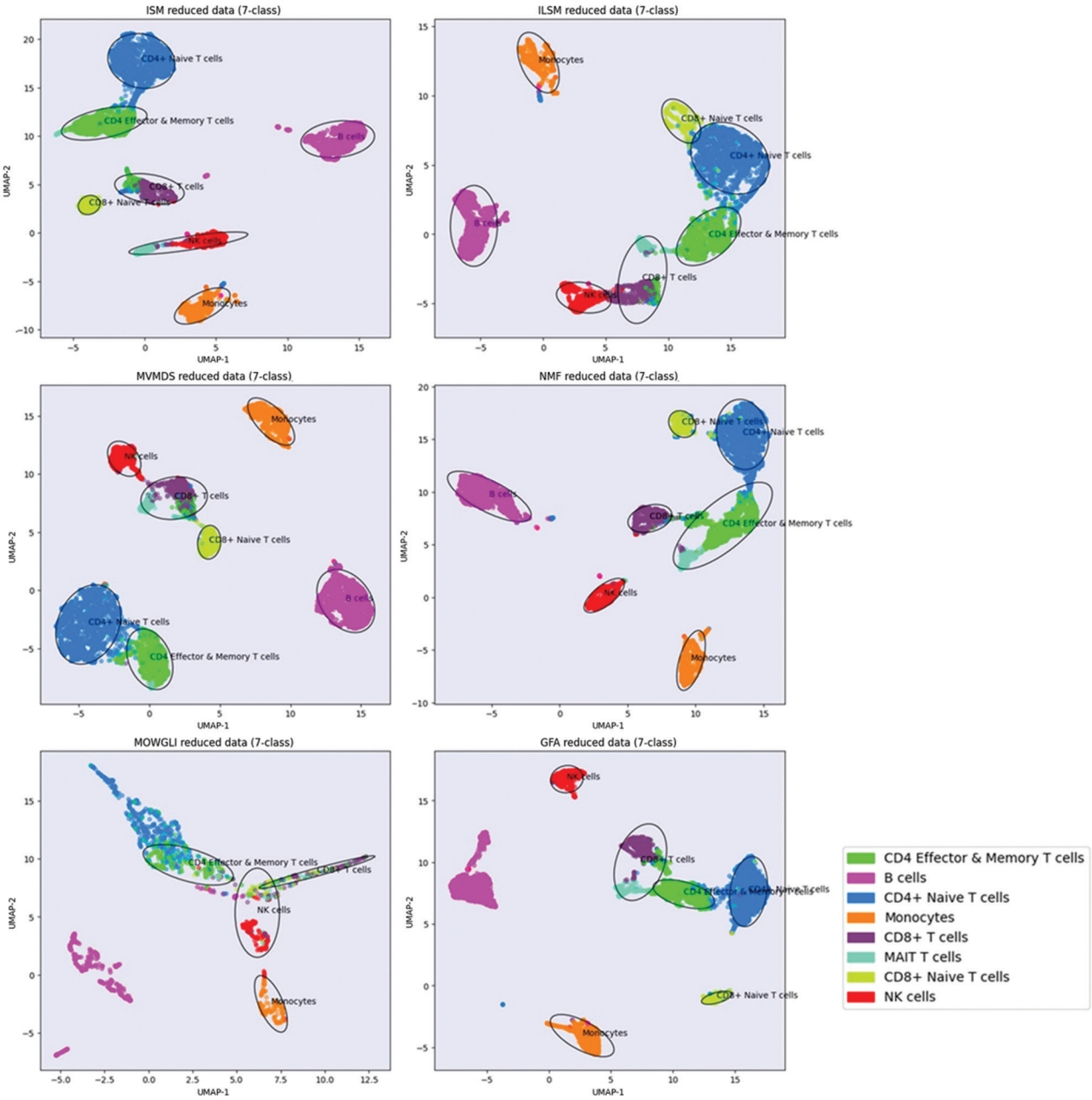


Figure 8. TEA-seq multi-omic single-cell data: Clustering of cells along ISM, ILSM, MVMDS, NMF, MOWGLI, and GFA components in the 2D scatterplots of the UMAP projection of the transformed data, with an additional cell type CD8+ naïve T cells as identified by ISM.
Abbreviations: GFA: Group factor analysis; ILSM: Integrated latent source model; ISM: Integrated sources model; MAIT: Mucosal-associated invariant T cell; MOWGLI: Multi-Omics Wasserstein inteGrative anaLysIs; MVMDS: Multi-view multidimensional scaling; NK: Natural killer cell; NMF: Non-negative matrix factorization; UMAP: Uniform manifold approximation and projection.

Several key observations are summarized as follows:

(i) The use of a sparsity parameter slightly >1 (e.g., 1.1) severely degrades performance across all metrics due to increased relative error and can even lead to computational errors, as observed in the UCI Digits dataset. Therefore, using a sparsity parameter too close to 1 poses a significant risk. To err on the conservative side, we choose the default value of 0.8.

(ii) The proportion of classes retrieved is not significantly affected by a low sparsity parameter. For example, eight out of 10-digit classes and 11 out of 16 cell types are still recognized with a sparsity parameter of 0.

Table 2. Relative error and other metrics as functions of the embedding dimension and rank (UCI Digits dataset, 10 classes)

Embedding (ISM only) rank used	Relative error	Proportion of classes retrieved	Purity	ARI	NMI	FMS	Sparsity	Specificity	Overall
(8,10)	0.84	0.60	0.27	0.38	0.50	0.44	0.86	0.24	0.47
(9,10)	0.52	1.00	0.58	0.57	0.67	0.61	0.87	0.20	0.64[a]
(10,10)	0.72	0.30	0.14	0.32	0.45	0.40	0.84	0.20	0.38
(11,10)	1.20	0.80	0.47	0.51	0.62	0.56	0.91	0.32	0.60
(12,10)	0.91	0.70	0.32	0.37	0.53	0.44	0.90	0.21	0.50
(9,8)	0.7	0.70	0.32	0.40	0.55	0.47	0.86	0.22	0.50
(9,9)	0.52	0.80	0.36	0.41	0.55	0.47	0.86	0.21	0.52
(9,10)	0.52	1.00	0.58	0.57	0.67	0.61	0.87	0.20	0.64[a]
(9,11)	0.62	0.90	0.42	0.44	0.58	0.50	0.87	0.19	0.56
(9,12)	0.61	1.00	0.62	0.60	0.67	0.64	0.88	0.21	0.66

Note: [a]The most performant combinations.
Abbreviations: ARI: Adjusted rand index; FMS: Fowlkes-Mallows score; ISM: Integrated sources model; NMI: Normalized mutual information index.

Table 3. Relative error and other metrics as functions of the embedding dimension and rank (Signature 915 dataset, 16 classes)

Embedding (ISM only) rank used	Relative error	Proportion of classes retrieved	Purity	ARI	NMI	FMS	Sparsity	Specificity	Overall
(14,16)	0.36	0.63	0.55	0.92	0.91	0.93	0.94	0.70	0.80
(15,16)	0.33	0.75	0.69	0.95	0.94	0.96	0.94	0.74	0.85
(16,16)	0.34	0.88	0.72	0.98	0.95	0.98	0.93	0.83	0.90[a]
(17,16)	0.34	0.81	0.67	0.92	0.91	0.93	0.94	0.74	0.85
(18,16)	0.38	0.81	0.75	0.95	0.94	0.96	0.94	0.74	0.87
(16,14)	0.39	0.69	0.60	0.96	0.94	0.97	0.92	0.76	0.83
(16,15)	0.34	0.88	0.73	0.98	0.96	0.98	0.93	0.84	0.90
(16,16)	0.34	0.88	0.72	0.98	0.95	0.98	0.93	0.83	0.90[a]
(16,17)	0.31	0.88	0.76	0.96	0.94	0.96	0.93	0.79	0.89
(16,18)	0.31	0.63	0.55	0.95	0.92	0.95	0.93	0.76	0.81

Note: [a]The most performant combinations.
Abbreviations: ARI: Adjusted rand index; FMS: Fowlkes-Mallows score; ISM: Integrated sources model; NMI: Normalized mutual information index.

Table 4. Effect of changing the sparsity parameter in the UCI Digits experiment (10 classes), with embedding dimension=9 and rank=10

Sparsity parameter	Relative error	Proportion of classes retrieved	Purity	ARI	NMI	FMS	Sparsity	Specificity	Overall
1.1	0.98	0.10	0.00	0.07	0.13	0.22	1.00	0.17	0.24
1	0.70	0.70	0.27	0.35	0.52	0.42	0.94	0.16	0.48
0.9	0.55	0.90	0.37	0.39	0.54	0.46	0.90	0.17	0.53
0.8	0.52	1.00	0.58	0.57	0.67	0.61	0.87	0.20	0.64
0.7	0.62	1.00	0.59	0.59	0.68	0.63	0.84	0.24	0.65
0.6	0.69	1.00	0.55	0.54	0.65	0.59	0.80	0.20	0.62
0.5	0.71	0.80	0.53	0.56	0.67	0.61	0.73	0.24	0.59
0.4	0.71	0.90	0.51	0.54	0.65	0.58	0.39	0.23	0.54
0.3	0.48	0.9	0.53	0.56	0.67	0.61	0.62	0.27	0.59
0.2	0.53	0.9	0.60	0.60	0.71	0.64	0.59	0.27	0.62
0.1	0.56	0.8	0.41	0.47	0.60	0.54	0.56	0.28	0.52
0	0.76	0.8	0.50	0.54	0.64	0.59	0.51	0.31	0.55

Abbreviations: ARI: Adjusted rand index; FMS: Fowlkes-Mallows score; NMI: Normalized mutual information index.

Table 5. Effect of changing the sparsity parameter in the Signature 915 experiment (16 classes) with embedding dimension=16 and rank=16

Sparsity parameter	Relative error	Proportion of classes retrieved	Purity	ARI	NMI	FMS	Sparsity	Specificity	Overall
1.1	-	-	-	-	-	-	-	-	-
1	0.36	0.81	0.70	0.98	0.95	0.98	0.93	0.80	0.88
0.9	0.34	0.88	0.71	0.98	0.95	0.98	0.93	0.83	0.89
0.8	0.34	0.88	0.72	0.98	0.95	0.98	0.93	0.83	0.90
0.7	0.34	0.63	0.50	0.96	0.92	0.96	0.93	0.83	0.82
0.6	0.33	0.81	0.67	0.97	0.93	0.97	0.93	0.79	0.87
0.5	0.34	0.63	0.54	0.96	0.92	0.96	0.92	0.83	0.82
0.4	0.34	0.69	0.61	0.96	0.94	0.97	0.93	0.72	0.83
0.3	0.33	0.75	0.65	0.97	0.94	0.97	0.93	0.72	0.85
0.2	0.34	0.75	0.67	0.95	0.94	0.96	0.93	0.72	0.85
0.1	0.33	0.81	0.70	0.94	0.93	0.95	0.92	0.72	0.85
0	0.31	0.69	0.61	0.91	0.90	0.92	0.92	0.74	0.81

Abbreviations: ARI: Adjusted rand index; FMS: Fowlkes-Mallows score; NMI: Normalized mutual information index.

This is due to the inherently high percentage of zero loadings in H when running standard NMF (on an average of 51% for the UCI Digits dataset and 92% for the Signature 915 dataset).

(iii) No metric shows an advantage of running ISM with a low sparsity parameter. For example, with the default sparsity parameter, higher sparsity and higher factor specificity are observed in the UCI Digits and Signature 915 experiments, respectively. To ensure a sufficient percentage of zero loadings regardless of the dataset, we recommend maintaining the default value of 0.8.

3.3.3. Evolution of the relative error over the course of model training

In this section, we evaluate how each main factorization step performed in the ISM workflow contributes to the final approximation error. Specifically, we examine the relative error obtained after (i) the preliminary NMF, (ii) the first call to NTF before the straightening process, and (iii) the last iteration of NTF in the straightening process.

While the increase in relative error is very small for the Signature 915 dataset (0.35 vs. 0.30), we observe a large increase for the UCI Digits dataset (0.53 vs. 0.36). This increase is mainly due to the straightening process (0.53 vs. 0.39 before). Recall that this process iteratively parsimonizes the view-mapping matrix H. The highly sparse nature of the Signature 915 dataset explains the difference in behavior between the two datasets: for the denser UCI Digits dataset, the increased sparsity of the view-mapping matrix induced by the straightening process significantly inflates the relative error, as more of the smaller values in the original views are filtered out. Unless the zero attribute loadings in some of the ISM components are relevant to digit class identification, this is not an issue. In fact, if we bypass the straightening process to achieve a smaller relative error, the performance of ISM is reduced; only nine-digit classes are found instead of 10, and the purity is 0.18 instead of 0.17, indicating that the model becomes overfit. This illustrates how ISM manages to filter out the specific part of the signal that is irrelevant to the main mechanisms in the data and hinders their recovery.

3.3.4. Computational time

In this section, we discuss the computation time required to analyze the TEA-seq dataset, which is a very large dataset (Table 6). The processing times for NMF, ISM, ILSM, and MVMDS are relatively short (0.55, 1.17, 1.31, and 5.31 min, respectively, on a computer equipped with an 11th Gen Intel® Core™ i7 processor and 16 GB of RAM, without the GPU activation). In contrast, GFA and MOFA+ require about 20 min with the GPU activated (7.9 GB available). MOWGLI is extremely slow, even with the GPU activated. For this reason, we had to consider only a 20% random sample of the Reuters and TEA-seq multi-omic single-cell datasets.

4. Discussion

The performance metrics used for our proof-of-concept analysis demonstrate that ISM performs as well as or better than other methods. The ISM workflow uses algorithms with proven performance and convergence properties, such as NMF and NTF, which is consistent with the good performance of ISM observed in our examples. In addition, the low computational time for large datasets indicates that this approach is highly scalable.

Table 6. Computational time observed in the TEA-seq multi-omic single-cell data

Method	Time (min)
MVMDS	5.31
ISM	1.17
ILSM	1.31
GFA	23.14
MOFA+	19.38
MOWGLI (20%)	82.31
NMF	0.55

Note: The parallelization of separate factorizations was not activated for ILSM, hence the slightly higher computational time compared to ISM. Abbreviations: GFA: Group factor analysis; ILSM: Integrated latent sources model; ISM: Integrated sources model; MOFA+: Multi-Omics Factor Analysis+; MOWGLI: Multi-Omics Wasserstein inteGrative anaLysIs; MVMDS: Multi-view multidimensional scaling; NMF: Non-negative matrix factorization.

However, the proportion of known categories retrieved and other metrics depend on the data being analyzed. For example, for the Reuters data, only three out of six categories are recognized at best using ISM or MVMDS, suggesting that latent-space-based methods may not be the most effective approaches with bag-of-words data.

In contrast to the other approaches studied, MVMDS and ISM are the only approaches that perform relatively well on all the datasets analyzed, demonstrating their versatility. The main advantages of ISM over MVMDS are its speed and increased sparsity in the latent-space representation. Regarding missing data, the ISM implementation uses an NTF package that can handle missing data, unlike MVMDS.

To the best of our knowledge, ISM is the first approach that uses NMF to transform heterogeneous views into a 3D array and then uses NTF to extract consistent information from the transformed views. However, apparent commonalities with anchor-based MVC methods (A-MVC) are worth mentioning to further illustrate the originality of ISM:

(i) In the first step, ISM relies on anchors, akin to A-MVC. ISM anchors correspond to zero-loading attributes in the latent spaces defined by the H_v, whereas A-MVC anchors are observations well distributed over existing clusters. Both act as intermediaries to derive either a latent space or cluster labels shared by all views.

(ii) In the second step, ISM applies NTF on the embedded views. A-MVC applies NTF on a tensor of anchor graphs, albeit with added constraints that ensure orthogonality and consistency in the cluster labels across all views.

A-MVC requires a specialized algorithm to select the anchor points that are best distributed across clusters. Since clusters must be sufficiently populated with A-MVC anchors for the method to work, the number of anchors must be set higher than the number of clusters. In contrast, ISM attribute anchors are found automatically through the process of parsimonization. This process requires the setting of a sparsity parameter to relax the reciprocal of the HHI, which may otherwise lead to excessive sparsity. In the examples considered in this article, this value is experiment independent and is set to 0.8. Further reducing the sparsity parameter risks a lack of overlap between the simplicial cones, potentially rendering the tensor decomposition ineffective. Therefore, until more experience is gained with ISM, we do not recommend changing this parameter.

Just as NMF and NTF factors are more interpretable and meaningful due to the non-negativity of their loadings, ISM produces latent factors whose interpretation is greatly facilitated by the non-negativity and sparsity of the attribute loadings. This is illustrated by the example of the Signature 915 dataset. It is noteworthy that all non-negative approaches result in a high sparsity index of the view-mapping, in contrast to the mixed-sign approaches.

The ISM has only three hyperparameters, which are very few compared to alternative methods: The sparsity coefficient, the embedding dimension, and the rank dimension. As mentioned, the sparsity coefficient should be kept at its default value of 0.8. Regarding the rank and embedding dimensions chosen for the ISM model, an objective and natural choice was the known number of classes for our examples, as we expect each factor to be distinctly assigned to a particular class. The only exception was the UCI digit dataset, where reducing the embedding dimension by one unit significantly reduced the error rate. However, this is only possible in a supervised setting where classes are known. More generally, as with all factorization methods, the factorization rank must be determined in advance.

This raises the issue of the subjectivity of the choice made, especially in an unsupervised setting where cross-validation cannot be used. For PCA, MVMDS, MOFA+, and GFA, setting the rank by inspecting the scree plot of the variance ratio is indeed a subjective choice due to the variety of possible criteria that can be used to identify an "elbow" in the scree plot. We tried a range of values around the "observed" elbow. The observed changes in the close neighborhood metric had no impact on the conclusions about the performance of ISM relative to other approaches (Tables S1 and S2). Since GFA and MOFA+ include automatic rank detection (ARD), increasing the rank should not adversely affect performance, as

it can be automatically reduced if the ARD criteria are met. Notably, for both experiments, increasing the chosen rank decreased performance in terms of cluster association with known classes. This again illustrates the difficulty of choosing the "right" rank. However, non-negative factorization-based methods, including ISM, are not subject to orthogonality constraints and can, therefore, create a new dimension by, for example, splitting a given component into two parts to disentangle close mechanisms that are otherwise intertwined in that component.[40] For this reason, the rank could be set to the number of known classes in a more logical and objective way. Finding the correct rank is, therefore, less critical than with mixed signed factorization approaches such as singular value decomposition (SVD), where low-variance components tend to represent the noisy part of the data. However, multiple solutions have been proposed, among which the cophenetic correlation coefficient is widely used to estimate a rank that provides the most stable clustering derived from the NMF components.[41] A similar criterion, named concordance, has been proposed,[31] where extensive simulations showed that NMF finds the most stable solutions around the correct rank, even if the latent factors are strongly correlated. While such an approach could be used with ISM to determine the best combination for the preliminary embedding and latent space dimensions, it would become too computationally intensive. However, in line with the fact that embedding and latent spaces are later merged in the ISM workflow, it can still be applied in the case where the model imposes the same dimensions for both parameters. As demonstrated in the proof-of-concept analysis of our examples, the embedding dimension can be further optimized by examining the approximation error in the neighborhood of the chosen rank.

Redundancy in the latent factors is a known issue for NMF-based techniques, as identified and illustrated early on with Donoho's swimmer dataset, where a ghost torso appeared in all basis vectors representing body parts in different orientations.[32] L1 regularization techniques, such as using Hoyer's sparsity index[42,43] or appropriate initialization like non-negative SVD (NNSVD),[44] can help mitigate these problems. Notably, in our ISM workflow implementation, the HHI used in the embedding step is mathematically equivalent to Hoyer's sparsity index, and NNSVD is used for NMF and NTF initialization.

ISM's intrinsic view loadings also enable the automatic weighting of views within each latent factor. This allows the simultaneous analysis of views of very different sizes without the need for prior normalization to give each view the same importance, as is necessary with methods like consensus PCA. However, this property reaches its limits when view sizes are extremely unbalanced, as seen in the prokaryotic dataset. In such cases, it is recommended to use ILSM, as ISM is applied to transformed views of equal size, giving equal weight to the original views with the smallest size, whereas global factorization tends to ignore them at initialization. In addition, ILSM requires significantly less computational time due to parallelizable view factorizations.

Recently, graph transformers and deep learning approaches have been proposed for the inference of biological single-cell networks.[45] The preliminary NMF in Unit 1 of Workflow 1, which combines the data before the application of NTF, is somewhat reminiscent of the "attention" mechanism used in transformers before the application of a lightweight neural network.[46] This could explain why ISM can outperform NTF when applied to a multidimensional array, even if the data structure is suitable for the direct application of NTF, as shown by the clustering of marker genes achieved in the Signature 915 dataset example. This also explains why, in the first two examples, although NMF is close to ISM in terms of purity index and other metrics, ISM outperforms NMF in terms of the number of classes detected and, in the second example, by generating a better positioning of the detected cell types on the 2D map projection. Likewise, in the multi-omic single-cell TEA-seq dataset, only ISM identifies and places a naïve cell subtype next to the most biologically relevant one.

Like other latent space methods, ISM is not limited to the purpose of MVC. The ISM components and the view-mapping matrix can be used for data reduction on newly collected data (i.e., data that is not part of the data used to train/learn the model) by fixing these components in the ISM model. Data reduction for newly collected data remains feasible even if some of the views contained in the training data are missing, as the ISM parameters are compartmentalized by views.

The ISM is not limited to views with non-negative data. Each mixed-signed view can be split into its positive part and the absolute value of its negative part, resulting in two different non-negative views, as illustrated in the UCI Digits and prokaryotic data examples.

An important limitation of ISM and other multi-view latent space approaches is the requirement for the availability of multi-view data for all observations in the training set. For financial or logistical reasons, a particular view may be missing in a subset of the observations, and this subset may vary depending on the view under consideration. We are currently developing a variant of ISM that can process multi-view data with missing views.

In this approach, sets of views with enough common observations are integrated with ISM separately. Using the model parameters, the transformation into the latent ISM space can be expanded to all views over all observations in the set, resulting in much larger transformed views than the original intersection would allow. This expansion process enables the integration of the ISM-transformed data from the different view sets, again using the ISM. Interestingly, a similar integrated latent space approach has already been proposed to study the influence of social networks on human behavior.[47] After masking a large number of views, the dataset of UCI Digits dataset was analyzed using this approach. A more detailed description of the expansion process (Workflow S1, Figure S1) and preliminary results (Figure S2) can be found in the Supplementary Materials.

Important issues such as the handling of highly dynamic or rapidly updating datasets have not yet been investigated. This will be addressed in a future article.

It is worth noting that by replacing NMF with NTF in the initialization unit of the ISM workflow, ISM can be easily extended to multi-view data where the views are themselves tensors of order three or higher, provided that all dimensions except the attribute dimension are shared between the views. Interesting applications include the analysis of longitudinal multi-view data or the integration of multiple X-ray views. These topics will be the subject of dedicated articles.

Finally, the extension of ISM to the ILSM approach, as described in the methods section (Section 2), is achieved by a simple chained matrix multiplication – an example of ISM inheriting the simplicity and compactness of the NTF model, made possible by embedding views in a 3D array. This has important advantages:

i. Performance
   - Independent view factorizations can be achieved using parallel computing.
   - The number of attributes in each transformed view is reduced to its factorization rank, allowing ISM to be performed on a much smaller dataset.

ii. Versatility
   - ILSM can be applied to compute NMF on big data in a federated or distributed way. To this end, smaller slices are constructed at random, with each slice considered a particular view that is submitted to ISM. Preliminary results indicate significant performance improvements (Workflow S2 and example in the Supplementary Materials).

While ILSM does not claim to outperform all alternative approaches in every context, this illustrates the scalability and versatility of ILSM, extending far beyond the scope of multi-view data analysis.

5. Conclusion

The proof-of-concept analysis results provide strong preliminary support for the proposed new method. As a next step, we will perform a comprehensive comparison of ISM with state-of-the-art alternative methods, including those considered in this article, and report the findings in a follow-up article.

To further illustrate ISM's key benefits and broad applicability, we will conclude by presenting some potential applications currently under evaluation, with results to be published in future articles.

In longitudinal clinical studies, where participants are followed up later, the ISM model can be trained at baseline and applied to subsequent data to calculate meta-scores. The interpretability of the associated components makes ISM meta-scores more appealing to clinicians compared to the mixed-sign latent factors from other factorization methods.

Consider complex multidimensional multi-omics data from one and the same set of cells (single-cell technology). There is a growing amount of single-cell data corresponding to different molecular layers of the same cell. Data integration is a challenge as each modality can provide a different clustering stemming from a specific biological signal. Therefore, data integration and its projection into a space must: (i) preserve the consensus between two clusterings and (ii) highlight the differences each modality may bring. ISM view loadings can address these two key requirements: components with similar contributions from each molecular layer highlight a consensus that can be inferred from clustering based on the ISM meta-scores of such components. In contrast, components with differing contributions from each molecular layer highlight each modality's specificities, which can be inferred from clustering based on the ISM meta-scores of such components.

The area of spatial mapping, including spatial imaging and spatial transcriptomics, is expanding at an unprecedented pace. An effective method for integrating different levels of information, such as gene or protein expression and spatial organization of cell phenotypes, is an unmet methodological need. We believe that ISM can integrate these different levels of information, as shown in the analysis of the UCI Digits data, to capture the constituents that allow spatial patterns to be distinguished across all levels.

The identification of new chemotypes with biological activity similar to that of a known active

molecule is an important challenge in drug discovery, known as "scaffold hopping."[48] In this context, we are currently analyzing the fingerprints of the docking of 10 of 1000 of molecules to dozens of proteins, with protein-associated fingerprints forming the different views of each molecule. The goal is to use the ISM-transformed fingerprints to predict scaffold-hopping chemotypes. Given the enormous size of the dataset – each fingerprint contains more than 100 binary digits – the ILSM strategy is being evaluated as a possible way to reduce computational problems, as smaller sets of views can be analyzed on smaller subsets of observations before integrating them in their entirety.

Acknowledgments

Our sincere thanks to Prasad Chaskar, Translational Medicine Senior Expert Data Science Lead at Galderma, for stimulating discussions, especially on potential limitations arising from missing views when training latent models with multiple views. We also thank Philippe Pinel from the Center for Computation Biology, Mines Paris/PSL, and Iktos SAS, Paris France, for discussions on addressing ISM calculation challenges in Computational Biology.

Funding

None.

Conflict of interest

Franck Augé and Galina Boldina are employees of Sanofi and may hold shares and/or stock options in the company. All other authors declare no conflicts of interest.

Author contributions

Conceptualization: Paul Fogel, George Luta
Investigation: Franck Augé, Galina Boldina
Writing-original draft: Paul Fogel, Christophe Geissler, Galina Boldina
Writing-review & editing: George Luta, Christophe Geissler, Franck Augé

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The data used in this article and the ISM Jupyter Python notebook can be downloaded from the Advestis part of Mazars GitHub repository (https://github.com/Advestis/adilsm).

Further disclosure

The paper has been uploaded to or deposited in a preprint server (https://www.preprints.org/manuscript/202402.1001/v3).

References

1. Cichocki A, Zdunek R, Phan AH, Amari S. Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation. IEEE Signal Process Mag. 2009;25:142-145.

   doi: 10.1002/9780470747278

2. Perry R, Mischler G, Guo R, *et al*. mvlearn: Multiview machine learning in python. *J Mach Learn Res*. 2020;22(109):1-7.

   doi: 10.48550/arXiv.2005.11890

3. Argelaguet R, Velten B, Arnol D, *et al*. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*. 2018;14(6):e8124.

   doi: 10.15252/msb.20178124

4. Argelaguet R, Arnol D, Bredikhin D, *et al*. MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol*. 2020;21(1):111.

   doi: 10.1186/s13059-020-02015-1

5. Wu J, Lin Z, Zha H. Essential tensor learning for multi-view spectral clustering. *IEEE Trans Image Process*. 2019;28(12):5910-5922.

   doi: 10.1109/tip.2019.2916740

6. Guo W, Che H, Leung M. Tensor-based adaptive consensus graph learning for multi-view clustering. *IEEE Trans Consum Electron*. 2024.

   doi: 10.1109/tce.2024.3376397

7. Li J, Gao Q, Wang Q, Xia W, Gao X. Multi-View Clustering via Semi-Non-Negative Tensor Factorization. *arXiv* [Preprint]; 2023.

   doi: 10.48550/arXiv.2303.16748

8. Wang S, Cao J, Lei F, Jiang J, Dai Q, Ling BW. Multiple kernel-based anchor graph coupled low-rank tensor learning for incomplete multi-view clustering. *Appl Intell*. 2022;53(4):3687-3712.

   doi: 10.1007/s10489-022-03735-6

9. Zhao W, Gao Q, Li G, Deng C, Yang M. One-Step Multi-View Clustering Based on Transition Probability. *arXiv* [Preprint]; 2024.

   doi: 10.48550/arXiv.2403.01460

10. Ali W, Yang M, Ali M, Ud-Din S. Fuzzy model-based sparse clustering with multivariate t-mixtures. *Appl Artif Intell*. 2023;37(1):2169299.

    doi: 10.1080/08839514.2023.2169299

11. Yang M, Hussain I. Unsupervised multi-view k-means clustering algorithm. *IEEE Access*. 2023;11:13574-13593.

    doi: 10.1109/access.2023.3243133

12. Hussain I, Sinaga KP, Yang M. Unsupervised multiview fuzzy C-means clustering algorithm. *Electronics*. 2023;12(21):4467-4467.

    doi: 10.3390/electronics12214467

13. Smilde AK, Westerhuis JA, de Jong S. A framework for sequential multiblock component methods. *J Chemometr*. 2003;17(6):323-337.

    doi: 10.1002/cem.811

14. Trendafilov NT. Stepwise estimation of common principal components. Comput Stat Data Anal. 2010;54(12):3446-3457.

    doi: 10.1016/j.csda.2010.03.010

15. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. Eur J Oper Res. 2014;238(2):391-403.

    doi: 10.1016/j.ejor.2014.01.008

16. Zhang C, Hu Q, Fu H, Zhu PF, Cao X. Latent Multi-View Subspace Clustering. In: *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*); 2017. p. 4333-4341.

    doi: 10.1109/cvpr.2017.461

17. Chen M, Huang L, Wang C, Huang D. Multi-view clustering in latent embedding space. *Proc AAAI Conf Artif Intell*. 2020;34(4):3513-3520.

    doi: 10.1609/aaai.v34i04.5756

18. Leppäaho E, Ammad-ud-din M, Kaski S. GFA: Exploratory analysis of multiple data sources with group factor analysis. *J Mach Learn Res*. 2017;18(1):1294-1298.

19. Zhao S, Gao C, Mukherjee S, Engelhardt BE. Bayesian group factor analysis with structured sparsity. *J Mach Learn Res*. 2016;17(1):6868-6914.

20. Zhang X, Zhao L, Zong L, Liu X, Yu H. Multi-view Clustering via Multi-Manifold Regularized Nonnegative Matrix Factorization. In: *IEEE International Conference on Data Mining*; 2014. p. 1103-1108.

    doi: 10.1109/icdm.2014.19

21. Huizing G, Deutschmann IM, Peyré G, Cantini L. Paired single-cell multi-omics data integration with Mowgli. *Nat Commun*. 2023;14(1):7711.

    doi: 10.1038/s41467-023-43019-2

22. Brbic M, Kopriva I. Multi-view low-rank sparse subspace clustering. *Pattern Recognit*. 2018;73:247-258.

    doi: 10.1016/j.patcog.2017.08.024

23. Dong Y, Che H, Leung MF, Liu C, Yan Z. Centric graph regularized log-norm sparse non-negative matrix factorization for multi-view clustering. *Signal Process*. 2024;217:109341.

    doi: 10.1016/j.sigpro.2023.109341

24. Fu L, Lin P, Vasilakos AV, Wang S. An overview of recent multi-view clustering. *Neurocomputing*. 2020;402:148-161.

    doi: 10.1016/j.neucom.2020.02.104

25. Dua D, Graff C. *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science; 2017. Available from: https://archive.ics.uci.edu/dataset/72/multiple+features

26. Boldina G, Fogel P, Rocher C, Bettembourg C, Luta G, Augé F. A2Sign: Agnostic algorithms for signatures-a universal method for identifying molecular signatures from transcriptomic datasets prior to cell-type deconvolution. *Bioinformatics*. 2021;38(4):1015-1021.

    doi: 10.1093/bioinformatics/btab773

27. Lewis DD, Yang Y, Rose TG, Li F. RCV1: A new benchmark collection for text categorization research. *J Mach Learn Res*. 2004;5:361-397.

28. Brbic M, Piškorec M, Vidulin V, Kriško A, Šmuc T, Supek F. The landscape of microbial phenotypic traits and associated genes. *Nucleic Acids Res*. 2016;44:10074-10090.

    doi: 10.1093/nar/gkw964

29. Swanson E, Lord C, Reading J, *et al*. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *Elife*. 2021;10:e63632.

    doi: 10.7554/eLife.63632

30. Hirschman AO. The paternity of an index. *Am Econ Rev*. 1964;54:761-762.

31. Fogel P, Geissler C, Morizet N, Luta G. On rank selection in non-negative matrix factorization using concordance. *Mathematics*. 2023;11(22):4611.

    doi: 10.3390/math11224611

32. Badeau R, Bertin N, Vincent E. Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization. *IEEE Trans Neural Netw*. 2010;21(12):1869-1881.

    doi: 10.1109/tnn.2010.2076831

33. Donoho DL, Stodden V. When does non-negative matrix factorization give a correct decomposition into parts? *Adv Neural Inf Process Syst*. 2003;16:1141-1148.

    doi: 10.7916/d88d05n7

34. Hubert L, Arabie P. Comparing partitions. *J Classif*. 1985;2(1):193-218.

    doi: 10.1007/BF01908075

35. Strehl A, Ghosh J. Cluster ensembles-A knowledge reuse framework for combining multiple partitions. *J Mach Learn Res*. 2002;3:583-617.

doi: 10.1162/153244303321897735

36. Fowlkes EB, Mallows CL. A method for comparing two hierarchical clusterings: Rejoinder. *J Am Stat Assoc*. 1983;78(383):553-569.

doi: 10.2307/2288123

37. Demaine ED, Hesterberg A, Koehler F, Lynch J, Urschel JC. Multidimensional Scaling: Approximation and Complexity. In: *Proceedings of the 38th International Conference on Machine Learning*; 2021. p. 2568-2578.

doi: 10.48550/arXiv.2109.11505

38. Zhai Z, Lei YL, Wang R, Xie Y. Supervised capacity preserving mapping: A clustering guided visualization method for scRNA-seq Data. *Bioinformatics*. 2022;38(9):2496-2503.

doi: 10.1093/bioinformatics/btac131

39. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine Learning in Python. *arXiv* [Preprint]; 2011.

doi: 10.48550/arXiv.1201.0490

40. Fogel P, Hawkins DM, Beecher C, Luta G, Young SS. A tale of two matrix factorizations. *Am Stat*. 2013;67(4):207-218.

doi: 10.1080/00031305.2013.845607

41. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*. 2004;101(12):4164-4169.

doi: 10.1073/pnas.0308531101

42. Hoyer PO. Non-negative matrix factorization with sparseness constraints. *J Mach Learn Res*. 2004;5:1457-1469.

doi: 10.48550/arXiv.cs/0408058

43. Potluru VK, Plis S, Le Roux J, Pearlmutter BA, Calhoun VD, Hayes TP. Block Coordinate Descent for Sparse NMF. International Conference on Learning Representations (ICLR); 2013.

44. Boutsidis C, Gallopoulos E. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognit*. 2008;41(4):1350-1362.

doi: 10.1016/j.patcog.2007.09.010

45. Ma A, Wang X, Li J, *et al*. Single-cell biological network inference using a heterogeneous graph transformer. *Nat Commun*. 2023;14(1):964.

doi: 10.1038/s41467-023-36559-0

46. Vaswani A, Shazeer NM, Parmar N, *et al*. Attention is all you need. *Neural Inf Process Syst*. 2017;30:5998-6008.

47. Park J, Jin IH, Jeon M. How social networks influence human behavior: An integrated latent space approach for differential social influence. *Psychometrika*. 2023;88:1529-1555.

doi: 10.1007/s11336-023-09934-5

48. Pinel P, Guichaoua G, Najm M, *et al*. Exploring isofunctional molecules: Design of a benchmark and evaluation of prediction performance. *Mol Inform*. 2023;42(4):e2200216.

doi: 10.1002/minf.202200216

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Interpretability analysis of deep models for COVID-19 detection

Daniel Peixoto Pinto da Silva[1], Edresson Casanova[2], Lucas Rafael Stefanel Gris[3], Marcelo Matheus Gauy[4]*, Arnaldo Candido Junior[5], Marcelo Finger[4], Flaviane Romani Fernandes Svartman[6], Beatriz Raposo de Medeiros[7], Marcus Vinícius Moreira Martins[8], Sandra Maria Aluísio[2], Larissa Cristina Berti[9], and João Paulo Teixeira[10]

[1]Academic Department of Computing, Federal University of Technology – Paraná, Medianeira, Paraná, Brazil

[2]Department of Computer Science, Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, São Paulo, Brazil

[3]Institute of Informatics, Federal University of Goiás, Goiania, Goiás, Brazil

[4]Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, São Paulo, Brazil

[5]Department of Computing and Statistics, Institute of Biosciences, Humanities and Exact Sciences, São Paulo State University, São José do Rio Preto, São Paulo, Brazil

[6]Department of Classical and Vernacular Literature, Faculty of Philosophy, Language, Literature and Human Sciences, University of São Paulo, São Paulo, São Paulo, Brazil

[7]Department of Linguistics, Faculty of Philosophy, Language, Literature and Human Sciences, University of São Paulo, São Paulo, São Paulo, Brazil

[8]Department of Literature and Linguistics, University of the State of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[9]Department of Speech Therapy, Faculty of Philosophy and Sciences, São Paulo State University, Marília, São Paulo, Brazil

[10]Department of Eletronics, Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Bragança, Portugal

***Corresponding author:**
Marcelo Matheus Gauy
(marcelo.gauy@usp.br)

Citation: da Silva DPP, Casanova E, Gris LRS, *et al*. Interpretability analysis of deep models for COVID-19 detection. *Artif Intell Health*. 2024;1(3):114-126. doi: 10.36922/aih.2992

Received: February 21, 2024

Accepted: June 17, 2024

Published Online: July 30, 2024

Abstract

During the coronavirus disease 2019 (COVID-19) pandemic, various research disciplines collaborated to address the impacts of severe acute respiratory syndrome coronavirus-2 infections. This paper presents an interpretability analysis of a convolutional neural network-based model designed for COVID-19 detection using audio data. We explore the input features that play a crucial role in the model's decision-making process, including spectrograms, fundamental frequency (F0), F0 standard deviation, sex, and age. Subsequently, we examine the model's decision patterns by generating heat maps to visualize its focus during the decision-making process. Emphasizing an explainable artificial intelligence approach, our findings demonstrate that the examined models can make unbiased decisions even in the presence of noise in training set audios, provided appropriate preprocessing steps are undertaken. Our top-performing model achieves a detection accuracy of 94.44%. Our analysis indicates that the analyzed models prioritize high-energy areas in spectrograms during the decision process, particularly focusing on high-energy regions associated with prosodic domains, while also effectively utilizing F0 for COVID-19 detection.

Keywords: Coronavirus disease 2019 detection; Voice processing; Gradient-weight class activation mapping

1. Introduction

In December 2019, a novel coronavirus, namely severe acute respiratory syndrome coronavirus-2, was identified as the causative agent for coronavirus disease 2019 (COVID-19). This coronavirus variant rapidly became a global concern, reaching pandemic status as declared by the World Health Organization.[1] COVID-19 evolved to become more contagious and lethal over a short period.

Researchers from all fields joined efforts to tackle the pandemic crisis. In particular, researchers in artificial intelligence (AI) and related areas sought methods to simplify COVID-19 detection. These methods use a variety of sources, such as medical examinations,[2] symptoms,[3] and X-ray images,[4] among others.[5] A potential source for COVID-19 detection is audio recordings. Several projects have collected audio samples from patients, including speech and cough sounds,[6-8] to develop detection models. These models could optimize the patient screening process. However, existing approaches have limitations in data collection procedures. For example, environmental noise can be present during audio capture, leading to model overfitting on such noise.

The dataset presented in the SPIRA project[8,9] illustrates these challenges. Positive audio samples (read speech) from COVID-19 patients were collected in hospitals, while samples from symptom-free individuals were obtained through a web interface. These samples were labeled as the control group, with the caveat that no additional testing for COVID-19 was performed on these subjects. Training a model on this dataset requires precautions to avoid learning biases due to differences in the collection environment, as patient audios may contain hospital noise, while the control group may include other environmental noises. Moreover, a model trained on this dataset contrasts healthy cases with more severe COVID-19 cases, which typically exhibit symptoms such as respiratory insufficiency. Such models will likely not be able to identify COVID-19 cases that do not induce severe symptoms.

In this work, we trained and analyzed convolutional neural networks (CNNs) for COVID-19 detection from audios using the dataset from the study by Casanova *et al*.[9] In addition, we analyzed factors important for the model decision using several criteria, namely spectrograms, fundamental frequency (F0), fundamental frequency standard deviation (F0-STD), speaker age, and sex. We applied the gradient-weight class activation mapping (Grad-CAM)[10] algorithm to generate heat maps, allowing us to investigate which pieces of information are most relevant for the model's classification decisions. As the dataset used in this work contained audios from different collection environments (hospital and domestic), learning biases could occur toward hospital noise. To mitigate this problem, we introduced hospital noise into domestic audio samples following the proposal of Casanova *et al*.[9] We also used their data augmentation techniques to improve model performance. Finally, we could literally hear the areas in the audio that the model values the most in its decision process. To achieve this, we multiplied the heat maps obtained from Grad-CAM by the original log-Mel spectrograms, and the result was synthesized. It is important to note that we focused on spectrograms rather than Mel-frequency cepstral coefficients (MFCCs)[11] to enhance interpretability, while previous works opted to explore MFCCs[9,12] to attain accuracies above or close to 90%. As spectrogram-based models were shown in those papers to have inherently lower accuracy, we employed methods such as transfer learning (e.g., pre-trained models on large-scale audio datasets) to recover the model's performance using log-Mel spectrograms.

As a result, our best model uses a pre-trained audio neural network (PANN)[13] called CNN14, which, through transfer learning, achieves 94.44% accuracy, in line with the accuracy on the same dataset from recent works[12] using transformers-based architectures.[14]

This work presents four main contributions:

(i) We present an analysis detailing the features crucial for deep models to detect or rule out COVID-19 in patient and control audios. In the analyzed data, spectrograms contain more discriminative information than the combination of F0, F0-STD, sex, and age. A visual analysis of heat maps generated by Grad-CAM shows that, among F0, F0-STD, sex, and age, the most important feature is F0.

(ii) We present an interpretation of the decisions made by deep models using heat maps and audio synthesis, following an explainable AI approach. Based on the heat maps and audio resynthesis, we formulate a few hypotheses for the factors affecting model decisions, such as (a) the structure of pauses (patients have longer and more frequent pauses than controls); (b) signal energy over time decreases faster for patients than controls; and (c) an interplay between syntax and prosody emerges as a boundary marked by formant vowel high energy.

(iii) Through manual analysis of the audio signals (using Grad-CAM), we ensure that the deep models focus on the voice (or silent pauses) rather than on environmental noise.

(iv) We demonstrate that models pre-trained on large-scale audio datasets, such as CNN14, can, through transfer learning, achieve accuracies on par with the best previously reported models,[12] even when using log-Mel spectrograms as input instead of MFCCs.

2. Related work

In the literature, COVID-19 detection has been studied using different types of input features for classification. From the perspective of feature analysis, these inputs can be roughly grouped into two categories: White-box or black-box, based on their ease of interpretation.

An example of an approach using mostly white-box features is the work of Bartl-Pokorny *et al.*[15] The authors used 88 features extracted from audios containing vowels to measure how COVID-19 patients differ from the control group. They found that F0-STD commonly varies between these two groups. In our work, we also used F0, F0-STD, and included sex and age as inputs for our deep models to detect COVID-19. Sex and age were included following the findings of previous works,[16-18] which identified that these factors influence F0 and F0-STD in COVID-19 patients. It was found that women and elderly subjects present more differences in these two parameters, as their voices become higher-pitched and less stable. Moreover, the study by Fernandes-Svartman *et al.*[18] demonstrated that the structure of pauses in speech undergoes significant changes between controls and hospitalized COVID-19 patients, even proposing a white-box model, which achieves above 87% accuracy using solely the speech pause distribution.

Regarding black-box features, Schuller *et al.*[19] proposed a challenge for COVID-19 detection from both speech and cough audios using the Cambridge COVID-19 Sound database.[6,7] They performed baseline experiments and identified thousands of features that can be used for general audio processing and, in particular, COVID-19 detection in audio. Zheng *et al.*[11] presented another example of black-box features, where MFCCs proved to be a useful method for COVID-19 detection while consuming few computational resources.[9] More robust approaches use spectrograms, transfer learning, and data augmentation for the task.[20] Recently, transformer-based architectures with MFCCs as input were used alongside transfer learning in the study by Gauy and Finger,[12] achieving accuracy above 95%. CNN-based PANNs (e.g., CNN14), which use spectrograms as input, were also used in a study by Gauy *et al.*,[21] achieving comparable accuracy to transformer-based architectures. Based on these results, we investigated the use of spectrograms for COVID-19 detection. In addition, transfer learning and data augmentation were employed in the study.

Related work either uses white-box features (e.g., sex and age) to better understand the effect of COVID-19 on patient's audio or black-box features (e.g., spectrograms) alongside deep learning for higher accuracy in COVID-19 detection tasks. In this work, we proposed an interpretability analysis of the decisions made by the deep learning models found in the literature, aiming for a better understanding of their results. To achieve this, we proposed the use of the Grad-CAM algorithm,[10] analyzing its heat maps and synthesizing audios based on those heat maps. These operations provide valuable insights into how deep models make their decisions. A similar approach to ours can be found in the study by Sobahi *et al.*,[22] where the authors used Grad-CAM to visualize the results generated by their proposed model for COVID-19 detection through cough sounds. Grad-CAM allowed them to identify which regions of the input were most relevant to the model's decision-making process. In addition, in a slightly different domain, previous works have used Grad-CAM to analyze COVID-19 detection models based on chest X-ray images.[23,24]

3. Methods

We used the SPIRA dataset from a previous study,[9] which contains spoken utterances from 432 speakers, including both patient and control group members. Audios were collected in COVID-19 wards where patients were hospitalized due to respiratory insufficiency, conventionally defined as a blood oxygen saturation level below 92%. Control group members were recorded using an application over the Internet. We used the same division into training, validation, and test sets as the previous study,[9] maintaining a balance by age and sex. Specifically, the dataset was divided into 292 training audios, 32 validation audios, and 108 test audios. The dataset includes recordings of patients and control group members speaking an utterance with no pre-defined pauses. The utterance is simple enough for most to understand but complex enough to present several polysyllabic words with primary and secondary stress syllables. The specific utterance was "*o amor ao próximo ajuda a enfrentar o coronavírus com a força que a gente precisa*" ("love for your neighbor helps face the coronavirus with the strength we need"). The dataset used is available at https://github.com/SPIRA-COVID19/SPIRA-ACL2021, and the codes for each model and experiment can also be found at https://github.com/danpeixoto/covid19-interpretability-analysis.

In this work, inspired by previous approaches,[12,13] we employed transfer learning methods from pre-trained models (described in Section 3.1). Following established methods,[20] we explored data augmentation techniques (described in Section 3.2). Similar to previous studies,[9,12] we utilized audio splitting based on windowing (described in Section 3.3). During training, we performed preprocessing steps (described in Section 3.4) on the audios, termed dynamic preprocessing,[9] to tackle overfitting issues. By combining all the aforementioned techniques, we performed six experiments, described in

Sections 3.5 and 3.6. The goal of these experiments was to determine which operations are relevant for classification and how they affect the model's decision process (analyzed by Grad-CAM).

The server used for our training has an Intel Xeon Silver CPU processor (39 cores, 2.40 GHz), 56 GB of RAM, and two Nvidia 2080 GPUs (8 GB of VRAM each). Some of the runs occurred only in the CPU cores, while others used both GPUs and the CPU. Overall, all the experiments took approximately the same time to run, around 6 h in the CPU-only scenario or 1 h using both GPUs and CPU. Some small variations were observed, mainly due to the preprocessing techniques used, as they were performed exclusively on the CPU during each epoch of training. It should be noted that inference took only a few seconds in our environment.

3.1. Transfer learning with PANNs

PANNs have proven effective for transfer learning across various tasks.[13] They have been successfully applied to multiple audio classification tasks, such as audio set tagging,[13] speech emotion recognition,[25] and automated audio captioning.[26] PANNs are Mel spectrogram-based models and trained on the AudioSet dataset, which comprises approximately 1.9 million audios, totalizing 527 classes and over 5000 h. While the original authors explored several architectures, in this work, we used only the CNN14 architecture due to its simplicity and similarity to SpiraNet,[9] allowing it to benefit from the same preprocessing techniques.

3.2. Data augmentation

Following the work of Casanova *et al.*,[20] three data augmentation techniques were applied: Noise insertion, Mix-up, and SpecAugment.

First, noise insertion was performed due to the different recording environments present in the SPIRA dataset for patient and control group audios. Previous research[9] has shown that models trained on this dataset can overfit if the data are not preprocessed adequately, leading to biases, such as distinguishing control and patient groups based on the presence of hospital ward noise. To mitigate this, we followed the Casanova *et al.*[9] approach by injecting hospital ward noise into all audios. For some experiments, we inserted four noise recordings for the control group and three for the patient group, while other experiments used three audio recordings for each class, based on Casanova *et al.*'s findings[9].

Second, due to the small size of the training set, we used a data augmentation technique called Mix-up to increase model robustness. Mix-up combines two random instances (*xi* and *xj*) from the training set and their respective classes (*yi* and *yj*) to generate a new instance[27] $\left(\tilde{x}, \tilde{y}\right)$ using Equations I and II:

$$\tilde{x} = \lambda x_i + \left(1-\lambda\right) x_j \quad \text{(I)}$$

$$\tilde{y} = \lambda y_i + \left(1-\lambda\right) y_j \quad \text{(II)}$$

where $\lambda \in [0, 1]$ is generated from a Beta distribution.

Unlike common image processing augmentation techniques, such as rotation, cropping, and horizontal flipping, Mix-up is applicable to various tasks, including audio processing.[28] It helps generate better decision frontiers in the manifold extracted by the model during training, which is particularly beneficial for small datasets.

Finally, to further enhance model robustness in the cases of small training sets, we used the SpecAugment[29] data augmentation technique. SpecAugment is designed for spectrograms and was initially developed for automatic voice recognition. It performs augmentation on the spectrogram by first applying distortion in the time dimension, termed time warping in the study by Casanova *et al.*,[9] and then masking parts of frequency channels and masking blocks in time. The frequency mask is applied over f consecutive Mel channels $[f0, f0 + f)$, where f is chosen uniformly from 0 to F and F represents the maximum number of masked Mel channels (set to eight in our experiments). The parameter $f0$ is uniformly chosen at random from $[0, v - f)$, where v is the total number of Mel-frequency channels. The temporal mask is performed over t time slots $[t0, t0 + t)$, where t and $t0$ are determined analogously to the frequency mask.

3.3. Windowing

Each original audio, which is at least 4 s long, is divided into smaller 4-s audios. The division was performed using a 4-s window with a 1-s hop. For example, a 5-s audio was split into two segments: The first from seconds 0 – 4 and the second from seconds 1 – 5. This approach, initially employed by Casanova *et al.*,[9] ensures uniform audio lengths and prevents model overfitting based on audio lengths. As patient audios tend to be longer, models can overfit on audio length if no normalization is done.

The windows cover repeated fragments of the original audios to include as many fragments of the original spoken sentence as possible. It is important to note that windowing was performed separately for training and test sets. During training, each fragment was labeled with the class of the original audio. In the test set, a voting mechanism over the windowed audios was used to determine the class

of original audio, as described by Casanova *et al.*[9] The voting summed the predicted probabilities for each class. Windowing also served as a simple data augmentation technique, in addition to the approaches presented in Section 3.2.

3.4. Dynamic preprocessing

The audios were preprocessed for each training step, ensuring a richer variety of augmented data. To maintain our model consistent, the same preprocessing was applied during the validation and test phases. The following operations were carried out:

(i) Noise injection
(ii) Windowing
(iii) Spectrogram extraction
(iv) Spec-augment application (only for training)
(v) Mix-up application (only for training)
(vi) Training step/test step.

Operations 4 and 5 were applied only to PANN-based experiments and only during training, while the other operations were common to all experiments. For operation 3, we used different parameters for spectrogram extraction in our experiments. Table 1 presents the two settings used across the experiments presented in Sections 3.5 and 3.6: Set 1 was used for SpiraNet and matched the parameters from Casanova *et al.*[9] and Set 2 was used for CNN4 and needed to be consistent with the parameters used in pre-training. Two parameters were common for all spectrogram-based experiments: The number of fast Fourier fransform[30] components (1200) and the spectrogram format (log-Mel).

3.5. Experiments to find the best inputs

Here, we describe three experiments aiming to estimate the accuracy of the SpiraNet[9] with respect to three different input configurations. These experiments investigated the role of different information types (spectrogram, F0, F0-STD, age, and sex) in the model's decision process. Spectrograms are matrices, while F0 is a vector, and the remaining data are scalars. We converted all these data into matrices to facilitate visual analysis using Grad-CAM, described in the subsequent sections. The representation is shown in Figure 1. The input, in its full form, has 401 × 120 pixels, where the spectrogram occupies the top region (401 × 80). Age, F0-STD, and sex occupy 20 lines, while age and sex use 133 columns and F0-STD uses 135 columns. Age is represented by shades of gray, as it is a scalar value, and F0-STD is similarly represented. Sex is a binary value, with zero for males and one for females. F0 is represented in a "bar code" style, where each value in the original vector is repeated across an entire column in the generated matrix.

Table 1. Settings used in the experiments

Set	Hop size (ms)	Number of frequency	Number of Mel	Window length (ms)
1	160	601	80	400
2	320	513	64	1,024

Abbreviation: ms: Milliseconds.

Using the scheme presented in Figure 1, the first three proposed experiments are:

- Experiment 1: Uses only the spectrogram (401 × 80 pixels) as input
- Experiment 2: Uses F0, F0-STD, age, and sex (401 × 40 pixels) as input
- Experiment 3: Uses all input data present in Figure 1, including the spectrograms, F0, F0-STD, age, and sex (401 × 120 pixels).

All three experiments are based on the SpiraNet model and use the configurations from Set 1 of Table 1. Moreover, the general hyperparameters for all the experiments (including Experiments 4 and 5 in Section 3.6), based on Casanova *et al.*,[9] are as follows: Binary cross-entropy loss and the Adam optimizer.[31] Given that the focus is on studying the model's decision process rather than performance, the batch size is set at one, early stopping and a learning rate scheduler are not used, and the number of epochs is set to 1000 for all experiments. Despite these settings, CNN14 achieves accuracies close to the best models reported in the literature. We used a fixed learning rate of 0.001 and a weight decay of 0.01.

3.6. Experiments over the training process

We performed three additional experiments to analyze classification models with respect to potential changes during training, pre-training, and post-processing. The three experiments are described as follows:

- Experiment 4: The goal of this experiment is to determine how the accuracy of a classification model changes when using large-scale pre-trained models. To achieve this, it focuses on pre-training, exploring the use of transfer learning through a PANN model (CNN14). This experiment was configured using Set 2 from Table 1.

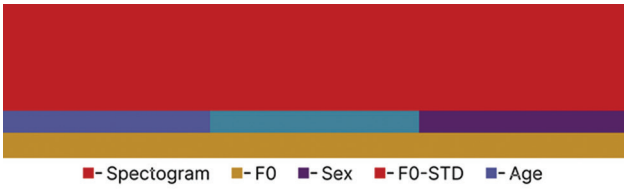


Figure 1. Input representation. Notes: F0: Fundamental frequency; F0-STD: Fundamental frequency standard deviation

- Experiment 5: This experiment aims to explore how the accuracy of a classification model changes when data augmentation techniques, such as SpecAugment and Mix-up, are used. In this experiment, we used the SpiraNet model but replaced MFCCs with spectrograms to simplify human analysis and improve audio resynthesis (see Experiment 6). As usual, for the SpiraNet, we use Set 1 of Table 1.
- Experiments 6a and 6b: These experiments differ from the previous five. We performed a qualitative analysis focused on model explainability using heat maps generated by Grad-CAM. This method aims to uncover the underlying reasons for the model's classification decisions by generating heat maps that highlight important zones in the decision process. First, in Experiment 6[a], we conducted a preliminary analysis and case study, investigating Grad-CAM from Experiments 1 to 3 (see Section 4.2) to understand which parts of the input are more relevant for classification. Then, in Experiment 6b, we performed a detailed analysis, focusing on the heat maps generated in Experiment 1. Our preliminary analysis showed that the spectrogram plays a major role in classification (see Section 4.3). In Section 4.3, we also resynthesized audios from Experiment 1, allowing us to hear them and investigate attention from both a visual and aural perspective. The audio reconstruction process is done in two steps. First, the heat map generated by Grad-CAM and the log-Mel spectrogram are combined using the Hadamard product. Second, the result and the phase of the original spectrogram are used to generate new audios highlighting the moments and frequencies the model considered most important in its decisions. We refer to the combination of original log-Mel spectrograms with heat maps as modified spectrograms.

4. Results

4.1. Experiments 1 – 5: Quantitative analysis

Table 2 presents the results of Experiments 1 – 5, with accuracies ranging from 65.75% to 94.44%. From Experiment 1, we observe that spectrograms are discriminative. Likewise, Experiment 2 showed that F0, F0-STD, sex, and age also contain discriminative information. However, spectrograms appear to carry more useful information since the accuracy of Experiment 1 is >10% higher than that of Experiment 2. Experiment 3 suggests that features extracted from inputs in Experiments 1 and 2 are largely equivalent despite a slight increase in accuracy (almost 2%) compared to Experiment 1. It should be noted that Experiments 1 - 4 used only noise insertion as data augmentation.

Table 2. Results from Experiments 1 – 5

Experiment	True		False		Accuracy (%)
	Positives	Negatives	Positives	Negatives	
1	37	49	5	17	79.63
2	36	38	16	18	68.52
3	44	44	10	10	81.48
4	51	51	3	3	94.44[a]
5	49	22	32	5	65.74

Note: [a]The highest accuracy among the experiments was achieved by Experiment 4.

Experiment 4 achieved the highest accuracy (94.44%), indicating that transfer learning significantly impacts learning features from patient and control groups, surpassing the results of Casanova *et al.*[9] These findings suggest that CNN14 might be better suited than SpiraNet for COVID-19 detection. CNN14's results are comparable in accuracy to those of transformers-based architectures described by Gauy and Finger,[12] with the added advantage of using spectrograms as input instead of MFCCs, as was the case for the MFCC-transformer.[12] This advantage is attributed to the effectiveness of the transfer learning used. Experiment 5 demonstrates that data augmentation techniques (SpecAugment and Mix-up) did not improve SpiraNet accuracy, as it performed worse than in Experiments 1 and 2. Experiments 6a and 6b are presented separately because they are based on heat maps, human analysis, and audio resynthesis (Sections 4.2 and 4.3, respectively).

Regarding errors, most experiments resulted in a balance of false positives and false negatives. Experiment 1 was an exception, presenting more false negatives. This experiment might have been more susceptible than others to cases of silent hypoxia, in which a patient has low blood oxygenation but does not present severe symptoms. Another exception was Experiment 5, which had significantly more false positives (32) than false negatives (5). A hypothesis for this phenomenon is that SpecAugment forces the model to give less importance to pauses, which are crucial for detecting respiratory insufficiency.[18] This may occur because the method introduces artificial pauses in training data.

4.2. Experiment 6a: Case study based on heat map analysis

Experiment 6a involved using Grad-CAM to generate heat maps for experiments based on inputs (Section 3.5). Figures 2-4 present the results of heat maps and modified spectrograms for Experiments 1 – 3, respectively.

Experiment 1 focused solely on spectrograms. The visual results for two patients and two control group

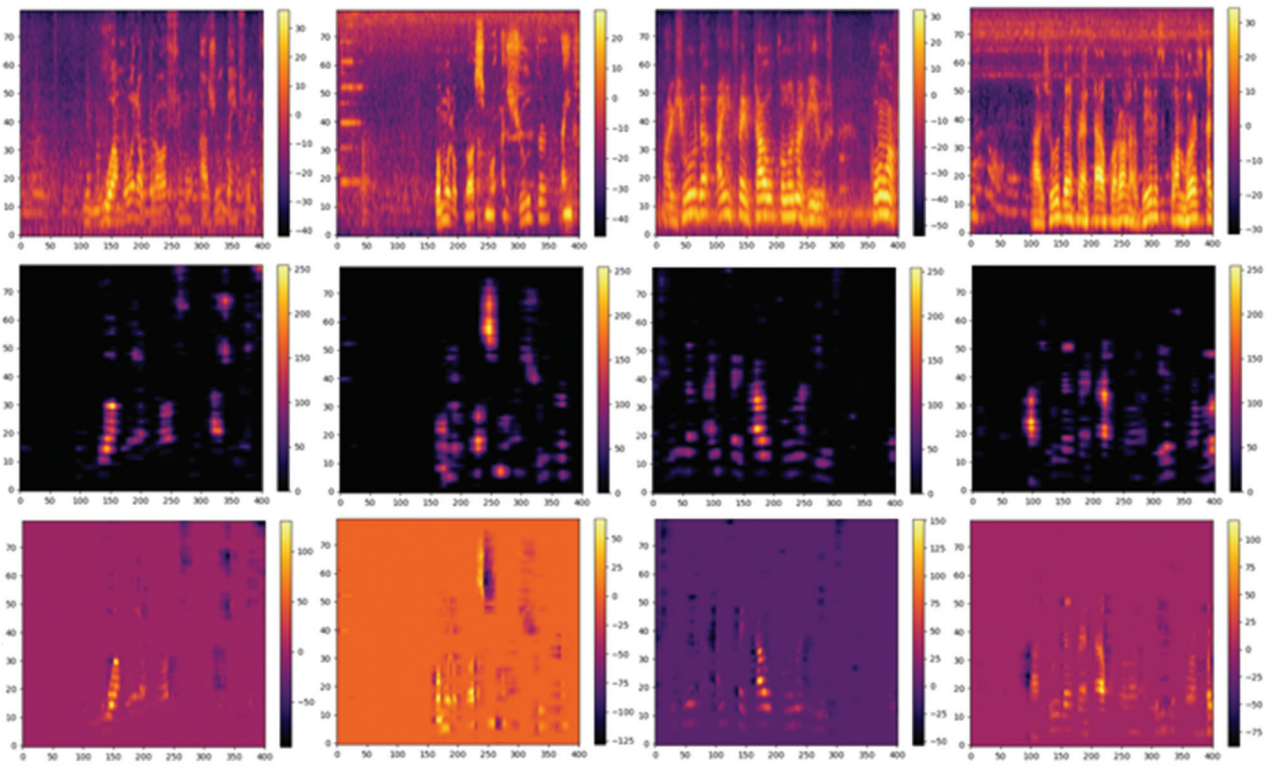


Figure 2. Results from Experiment 6a regarding Experiment 1 (spectrogram only), including original spectrograms (top), heat maps (middle), and modified spectrograms (bottom) for two control group members (left) and two patients (right).

members, including original spectrograms, heat maps, and modified spectrograms, are presented in Figure 2. We observe activity (attention) in high-energy regions over the input. These results suggest that energy levels and audio formats (*H*2 and *H*3) may play a significant role in COVID-19 detection.

In Section 4.1, the results indicated F0, F0-STD, age, and sex as distinctive features for COVID-19 detection. Figure 3 presents Experiment 2 visual representations for two patients and two control group members. It can be noted that F0 plays a major role in this model's detection process, especially in regions associated with transitions from voiced phonemes to pauses (*H*1) or to voiceless phonemes. The same applies to transitions from pauses or voiceless phonemes to voiced phonemes. In addition, sex and age appear to play a role in control classification, although not as noticeable as F0. On the other hand, F0-STD appears to be disregarded by the model.

Figure 4 presents the visual representations generated using all available information (spectrograms, F0, F0-STD, sex, and age). Heat maps suggest that spectrograms, F0, and sex are useful for patient classification, while control group detections are based only on spectrograms and F0. These observations indicate that spectrograms and F0 may contain complementary information, given the slightly superior accuracy obtained from Experiment 3 compared to Experiments 1 and 2.

4.3. Experiment 6b: Phonetical investigation and qualitative analysis

The phonetic investigation and qualitative analysis presented here were carried out by three linguists. Four main inputs were considered:

(i) Regular spectrograms in hertz were obtained from the original audios. These spectrograms were generated using the software PRAAT v6.1.09.
(ii) Original and modified Mel-spectrograms to highlight attention, as presented in Section 4.2 (Experiment 1).
(iii) Resynthesized audios from the modified spectrograms from the previous input allow us to hear where the model pays attention, while spectrograms show where the model focuses. These resynthesized audios are publicly available (https://drive.google.com/drive/folders/1aQEq82iUpnAmrQzQ52458GORv8PEK3nr?usp=share_link).
(iv) Regular spectrograms in hertz from audios resynthesized from our modified spectrograms obtained from the previous input. These spectrograms combine speech with heat maps.

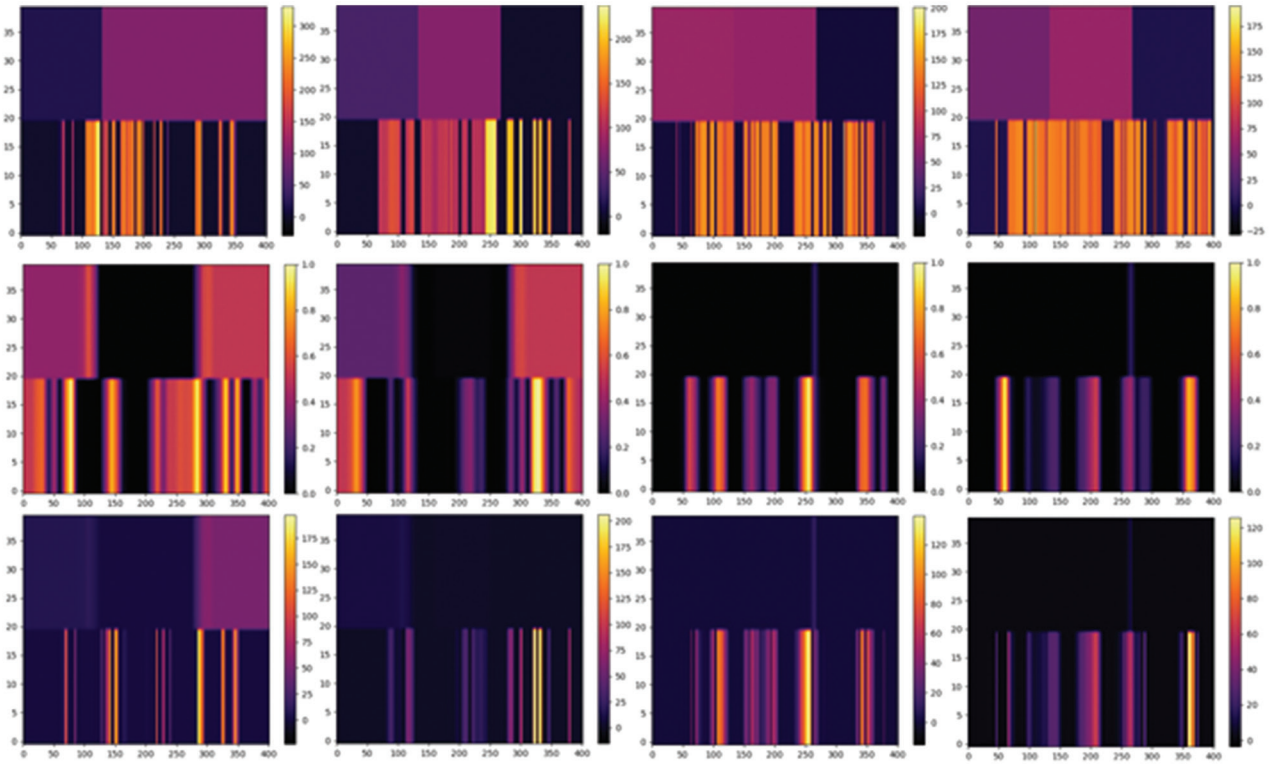


Figure 3. Results from Experiment 6a regarding Experiment 2 (all inputs except spectrograms), including original images (top), heat maps (middle), and modified images (bottom) for two control group members (left) and two patients (right).

Given our windowing approach, which involved generating 4-s windows with a 1-s hop, this analysis considered only the central audio fragment in the window. In total, the central fragments of 73 audios were inspected, containing 30 correct predictions for each class and 13 errors, from seven patients and six controls. It is important to note that Experiment 6b resynthesis is based on the model trained in Experiment 1. We chose Experiment 1 rather than 4 for better comparability with the analysis performed in Section 4.2.

It was observed that the decision process usually hinges on two aspects of the speech sound signal: The continuity of the signal versus its interruption. Thus, the model appears to pay attention to an alternation between the continuity of speech sounds and their discontinuity, which, in terms of intonational analysis, are pauses inserted by speakers. This observation is in line with the findings of Fernandes-Svartman *et al.*,[18] which noted that patients' pauses are significantly longer than those of control subjects and, being more frequent, are inserted in more places throughout the utterance.

Considering short-term parameters, such as the most salient vowels for the model, it was observed that the vowels/a/from "*ajuda a*" ("helps to") and "*enfrentar*" ("to face");/ɔ/from "*próximo*" ("neighbor");/o/from "*força*" ("strength");/oN/from "*com*" ("with") are those reproduced with more intensity in the modified spectrogram. This pattern corresponds to what occurs in the original audio spectrogram. Besides being expected, it is reasonable, since these vowels occupy prominent places in the utterance or are intrinsically more intense, such as the low vowels/a/and/ɔ/. On the other hand, the mid-high, oral/o/, and nasal/oN/ vowels do not have the same sound amplitude as the low ones but occupy a prosodically highlighted place in the utterance.

Therefore, on the one hand, we have the phonetic features of vowels and, on the other hand, the prosodic feature interacting with morpho-syntactic and semantic-pragmatic levels. The interaction discussed here explains the emphasis on the verb (there is usually a peak of F0 in the verbal item in a statement). In semantic terms, emphasis is given to "*com a força*" ("with the strength"). The initial expression of the adverbial phrase "*com a força que a gente precisa*" ("with the strength we need") is often phrased as an intonational phrase in our data. The intonational phrase initial position is prominent in Portuguese.[32,33] In pragmatic terms, this adverbial phrase is new information that modalizes the meaning of the verb "to face" (it is necessary to face the virus with strength).

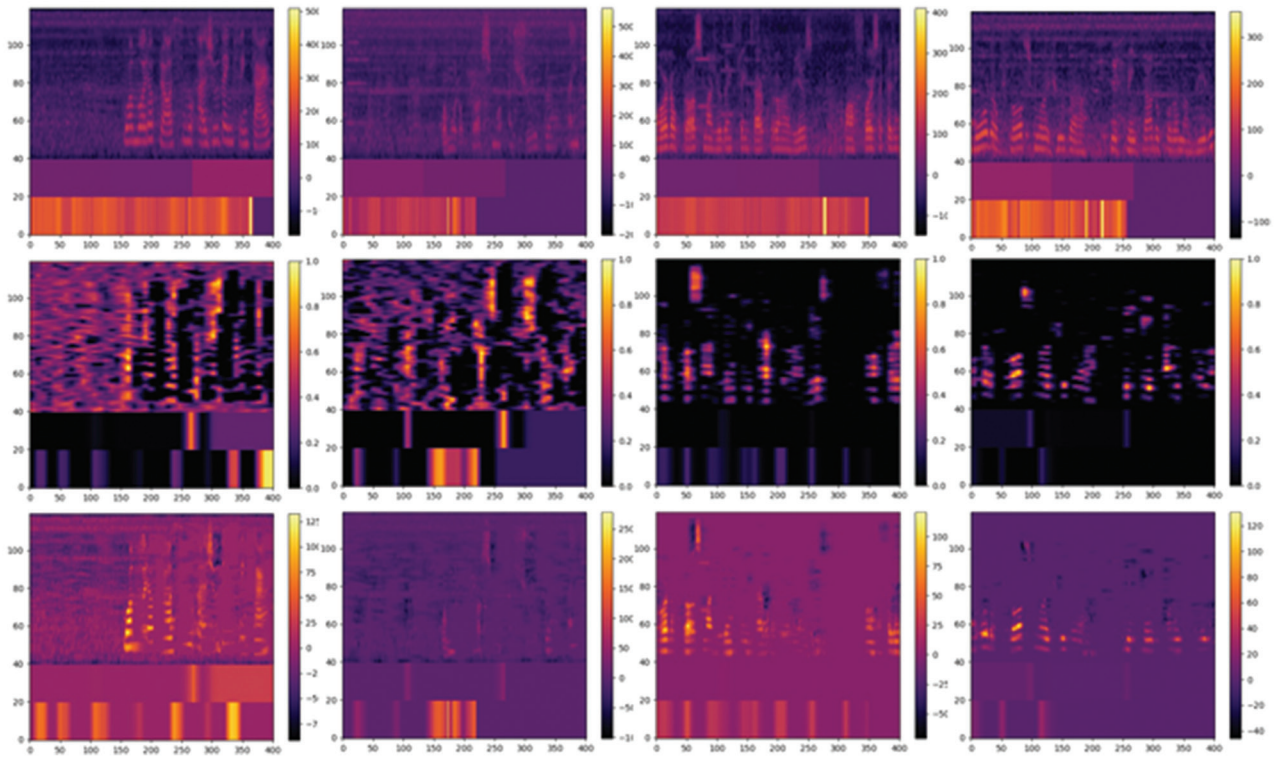


Figure 4. Results from Experiment 6a regarding Experiment 3 (all inputs), including original images (top), heat maps (middle), and modified images (bottom) for two control group members (left) and two patients (right).

Taking into account the phonetic analysis in interaction with other linguistic levels, the model shows the following preferences for distinguishing patients from controls:

i. Average patient pauses (approximately 400 ms) represent large interruptions of formant frequency tracks.
ii. Intrinsically more intense vowels are important clues. The model highlighted that the first (F1) and second (F2) formants in a 550 – 1300 Hertz range for both groups. For patients, these highlights can occur before or after a pause.
iii. The interaction between non-low vowels (with F1 and F2 ranging from 350 to 1000 Hz), the morpho-syntactic context, the prosodic domain in which they are produced, and their semantic-pragmatic role indicates that these segments are more prominent in the utterance, which appears important for the model.
iv. The interaction between non-low vowels and the initial position in the intonational phrase ("*com a força que a gente precisa*" ["with the strength we need"]) results in prosodic emphasis of this unit ("*com a força*" ["with the strength"]), which draws the model's attention.

Finally, for the correct predictions, regarding the interplay between the speech sound signal and the prosodic behavior that emphasizes it, we propose the following explanatory hypothesis: The model analyzes the signal as continuous and emphasizes some vowel formants in one of the groups; in another group, it focuses on important interruptions in a similar speech signal (the same sentence uttered by patients and controls). This approach leads to successfully distinguishing the two different speech groups, as patients with respiratory difficulties are unable to produce fluent speech and usually speak linguistic utterances with many pauses.

5. Discussion

In this section, we discuss a few hypotheses that can be deduced from our results (Section 5.1) as well as a few limitations of our approach and potential future work (Section 5.2).

5.1. Hypotheses for the model decision process

Regarding the question of which input features are best for the models, our results demonstrated that spectrograms convey important features for classification compared to other information, such as sex, gender, and F0-STD (Table 2). F0 also presented a small improvement during the classification process.

Regarding the training process, we found that noise insertion is important, consistent with previous findings;[20] therefore, we used it in all experiments. Other augmentations, such as Mix-up and SpecAugment, did not lead to improvements in the model. On the contrary, accuracy decreased. Transfer learning, on the other hand, proved to be important in this domain, as CNN14 achieved superior results compared to all other models and is comparable to the current state of the art in the literature for this task.

Furthermore, with respect to the training process, we noted in preliminary experiments some variance in the aspects a model can focus on during inference. The structure of pauses, syntactic boundaries, and pretonic syllables, among other factors, may be more or less evidenced by the models after training. This result is expected because artificial neural networks are high-variance, low-bias classifiers with randomized parameter initialization. We observed that transfer learning appears to reduce this variance.

Regarding the qualitative analysis, our first case study indicated that detailed evaluation would be better performed in the spectrograms-only scenario, which allowed for audio resynthesis, improving the process. As a result of this analysis, we can formulate the following hypotheses to explain the obtained variance and understand the data aspects that may play a role in model learning:

(i) H1: Pauses are important clues for detecting COVID-19 since patients tend to make more pauses for breathing than the control group.
(ii) H2: As the air starts decreasing in the lungs, the speaker may begin to lose breath, or the signal energy may begin to decrease. Thus, energy over time can be an important clue.
(iii) H3: An interplay between syntax and prosody is expected to emerge as a boundary marked by formant vowel high energy, i.e., phonetically.

The first hypothesis confirms that deep models use the discrepancy in the structure of pauses between patients and controls, as observed by Fernandes-Svartman *et al.*[18] The second and third hypotheses are newly observed discrepancies, which were found to be present by deep learning models.

Our work also confirms the hypothesis from previous works[9,12] that the addition of hospital ward noise, alongside suitable preprocessing steps, prevents the models from making biased decisions in the COVID-19 detection task. Through Grad-CAM analysis, we confirm that deep models focus on the voice (or silent pauses) rather than on environmental noise.

Finally, our best model (CNN14) achieved an accuracy of 94.44%. This number is almost as good as the best models reported in the literature[12] and shows that proper use of transfer learning can make log-Mel spectrogram input nearly as efficient as MFCC input.

5.2. Limitations and future work

In future works, we plan to investigate other audio-related features, such as autocorrelation, jitter, and shimmer. We also intend to investigate the beginning of a sentence. When a speaker starts to produce a sentence, they have more air in their lungs, which decreases as they speak. Some models may focus more on the audio at the beginning, measuring the signal energy, as the initial energy in the audio may provide hints about pulmonary capacity. In addition, we plan to investigate models of related diseases, such as general cases of respiratory insufficiency. Finally, we aim to investigate the variance in model training, identifying factors that are important for model inference and techniques that reduce variance in the learned models (such as transfer learning).

6. Conclusion

This work presents a method for interpretability analysis of audio classification for COVID-19 detection based on CNNs. Our work focuses on explainable AI. We investigated the importance of different features in the training process and generated heat maps to understand the model's reasoning for its predictions.

Regarding the input data, our results show that spectrograms are a suitable representation for COVID-19 detection. F0 appears to be almost as efficient as spectrograms, and the combination of these two inputs led to a small increase in the model performance. Grad-CAM analysis indicates that F0 is a more important feature than F0-STD, sex, and age. Moreover, Grad-CAM and audio resynthesis helped us formulate hypotheses about the factors that determine the model's classification process and confirm that the deep models used do not rely on environmental noise for decision-making. Our best model (CNN14) achieved 94.44% accuracy, on par with the best models in the literature12.

Acknowledgments

We gratefully acknowledge the support of NVIDIA corporation with the donation of a GPU used in part of the experiments presented in this research.

Funding

This work was supported by FAPESP grants 2022/16374-6 (MMG), 2020/06443-5 (SPIRA), and 2023/00488-5

(SPIRA-BM) and by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Arnaldo Candido Junior, Marcelo Finger
Investigation: Daniel Peixoto Pinto da Silva, Edresson Casanova, Arnaldo Candido Junior
Methodology: Daniel Peixoto Pinto da Silva, Lucas Rafael Stefanel Gris, Flaviane Romani Fernandes Svartman, Beatriz Raposo de Medeiros, Marcus Vinícius Moreira Martins, Larissa Cristina Berti
Writing – original draft: Daniel Peixoto Pinto da Silva, Arnaldo Candido Junior, Flaviane Romani Fernandes Svartman, Beatriz Raposo de Medeiros, Marcus Vinícius Moreira Martins, Larissa Cristina Berti
Writing – review & editing: Marcelo Matheus Gauy, Arnaldo Candido Junior, Sandra Maria Aluísio, João Paulo Teixeira, Marcelo Finger

Ethics approval and consent to participate

The research described in the paper was developed within the scope of the SPIRA Project (System for the Early Detection of Respiratory Insufficiency via Audio), which was approved by the Research Ethics Committee (IRB) of the Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HCFM/USP), Report 3.988.088, approved on April 24, 2020. The report states that this research does not require signed informed consent, as data collection involves voice assessment, and participants consent to participate by recording their acceptance on the equipment (cell phone) used in the study.

Consent for publication

Due to the pandemic, the IRB of the Hospital das Clínicas authorized us to collect patients' agreement to participate in the form of a recorded acceptance only. All participants expressed their agreement in a recorded audio.

Availability of data

The audio data can be found at https://github.com/SPIRA-COVID19/SPIRA-ACL2021/tree/master

Further disclosure

This paper has been uploaded to Arxiv at: https://arxiv.org/pdf/2211.14372.pdf. The code for the models can be found at: https://github.com/danpeixoto/covid19-interpretability-analysis.

References

1. *Who Director-General's Opening Remarks at the Media Briefing on Covid-19*. World Health Organization; 2020. Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19-11-march-2020 [Last accessed on 2024 Jul 19].

2. Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 infection from routine blood exams with machine learning: A feasibility study. *J Med Syst*. 2020;44(8):135.

   doi: 10.1007/s10916-020-01597-4

3. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digit Med*. 2021;4(1):3.

   doi: 10.1038/s41746-020-00372-6

4. 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

5. Acar E, Şahin E, Yılmaz İ. Improving effectiveness of different deep learning-based models for detecting COVID-19 from computed tomography (CT) images. *Neural Comput Appl*. 2021;33:17589-17609.

   doi: 10.1007/s00521-021-06344-5

6. Han J, Brown C, Chauhan J, *et al*. Exploring Automatic COVID-19 Diagnosis via Voice and Symptoms from Crowdsourced Data. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. p. 8328-8332.

   doi: 10.1109/ICASSP39728.2021.9414576

7. Brown C, Chauhan J, Grammenos A, *et al*. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD; 2020. p. 3474-3484.

   doi: 10.1145/3394486.3412865

8. Aluísio SM, Camargo Neto AC, Casanova E, *et al*. Detecting Respiratory Insufficiency via Voice Analysis: The SPIRA Project. In: *Practical Machine Learning for Developing Countries on the Tenth International Conference on Learning Representations*; 2022.

9. Casanova E, Gris L, Camargo A, *et al*. Deep learning against COVID-19: Respiratory insufficiency detection in Brazilian Portuguese speech. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP*; 2021. p. 625-633.

   doi: 10.18653/v1/2021.findings-acl.55

10. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual Explanations from Deep

Networks Via Gradient-Based Localization. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2017. p. 618-626.

doi: 10.1109/ICCV.2017.74

11. Zheng F, Zhang G, Song Z. Comparison of different implementations of MFCC. *J Comput Sci Technol*. 2001;16(6):582-589.

doi: 10.1007/BF02943243

12. Gauy MM, Finger M. Audio MFCC-Gram Transformers for Respiratory Insufficiency Detection in COVID-19. In: *Proceedings XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana,* STIL; 2021. p. 143-152.

doi: 10.5753/stil.2021.17793

13. Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley MD. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. Vol. 28. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*; 2020. p. 2880-2894.

doi: 10.1109/TASLP.2020.3030497

14. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. *Adv Neural Inform Process Syst*. 2017;30:5998-6008.

doi: 10.5555/3295222.3295349

15. Bartl-Pokorny KD, Pokorny FB, Batliner A, *et al*. The voice of COVID-19: Acoustic correlates of infection. *J Acoust Soc Am*. 2021;149(6):4377.

doi: 10.1121/10.0005194

16. Berti LC, Spazzapan EA, Pereira PL, *et al*. Mudanças Nos Parâmetros Acústicos da voz em Brasileiros com COVID-19. In: *XXIX Congresso Brasileiro e o IX Congresso Internacional de Fonoaudiologia*; 2021. p. 2819-2819.

17. Berti LC, Spazzapan EA, Queiroz M, *et al*. Fundamental frequency related parameters in Brazilians with COVID-19. *J Acoust Soc Am*. 2023;153:576-585.

doi: 10.1121/10.0016848

18. Fernandes-Svartman FR, Berti LC, Martins MVM, de Medeiros BR, Queiroz M. Temporal Prosodic Cues for COVID-19 in Brazilian Portuguese Speakers. In: *Proceedings Speech Prosody*; 2022. p. 210-214.

doi: 10.21437/SpeechProsody.2022-43

19. Schuller BW, Batliner A, Bergler C, *et al*. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation and Primates. In: *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH*; 2021.

doi: 10.21437/Interspeech.2021-19

20. Casanova E, Cândido A, Fernandes RC, *et al*. Transfer Learning and Data Augmentation Techniques to the COVID-19 Identification Tasks in COMPARE 2021. In: *22nd Annual Conference of the International Speech Communication Association*; 2021. p. 4301-4305.

doi: 10.21437/Interspeech.2021-1798

21. Gauy MM, Berti LC, Cândido Júnior A, *et al*. Discriminant Audio Properties in Deep Learning Based Respiratory Insufficiency Detection in Brazilian Portuguese. In: *Artificial Intelligence in Medicine: 21st International Conference on Artificial Intelligence in Medicine*; 2023. p. 271-275.

doi: 10.1007/978-3-031-34344-5_32

22. Sobahi N, Atila O, Deniz E, Sengur A, Acharya UR. Explainable COVID-19 detection using fractal dimension and vision Transformer with Grad-CAM on cough sounds. *Biocybern Biomed Eng*. 2022;42(3):1066-1080.

doi: 10.1016/j.bbe.2022.08.005

23. Moujahid H, Cherradi B, Al-Sarem M, *et al*. Combining CNN and Grad-CAM for COVID-19 disease prediction and visual explanation. *Intell Autom Soft Comput*. 2022;32(2):723-745.

doi: 10.32604/iasc.2022.022179

24. Panwar H, Gupta P, Siddiqui MK, Morales-Menendez R, Bhardwaj P, Singh V. A deep learning and Grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest x-ray and ct-scan images. *Chaos Solitons Fractals*. 2020;140:110190.

doi: 10.1016/j.chaos.2020.110190

25. Gauy MM, Finger M. Pretrained Audio Neural Networks for Speech Emotion Recognition in Portuguese. In: *First Workshop on Automatic Speech Recognition for Spontaneous and Prepared Speech Speech emotion recognition in Portuguese, SE&R*; 2022.

26. Xu X, Dinkel H, Wu M, Xie Z, Yu K. Investigating Local and Global Information for Automated Audio Captioning with Transfer Learning. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; 2021. p. 905-909.

doi: 10.1109/ICASSP39728.2021.9413982

27. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond Empirical Risk Minimization. In: *International Conference on Learning Representations*; 2018.

28. 28. Xu K, Feng D, Mi H, et al. Mixup-based acoustic scene classification using multi-channel convolutional neural network. In: Advances in Multimedia Information Processing. Vol. 11166. Cham: Springer; 2018. p. 14-23.

doi: 10.1007/978-3-030-00764-5_2

29. Park DS, Chan W, ZhangY, et al. SpecAugment: A simple data augmentation method for automatic speech recognition. Proc Interspeech. 2019;1:2613-2617.

doi: 10.21437/Interspeech.2019-2680

30. Brigham EO, Morrow R. The fast fourier transform. *IEEE Spectrum*. 1967;4(12):63-70.

doi: 10.1109/MSPEC.1967.5217220

31. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations*; 2015.

32. Frota S. *Prosody and Focus in European Portuguese: Phonological Phrasing and Intonation*. London: Routledge; 2014.

33. Tenani L. *Domínios Prosódicos no Português do Brasil: Implicações Para Prosódia e Para a Aplicação de Processos Fonológicos*. *Sínteses*; 2023. p. 8. Available from: https://revistas.iel.unicamp.br/index.php/sinteses/article/view/6275 [Last accessed on 2024 July 29].

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Experiences of Alzheimer's disease and related dementia family caregivers on Reddit communities: A topic modeling and sentiment analysis

Yulin Hswen[1,2], Jiangmei Xiong[3], Margaret Hurley[4], and Thu T. Nguyen[5]*

[1]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, United States of America

[2]Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, California, United States of America

[3]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

[4]John Snow Inc, Washington, D.C., United States of America

[5]Department of Epidemiology and Biostatistics, College Park, University of Maryland School of Public Health, District of Columbia, United States of America

Abstract

Alzheimer's disease and related dementias (ADRD) are a spectrum of disorders characterized by cognitive decline, which pose significant challenges for both affected individuals and their caregivers. Previous literature has focused on patient family surveys which do not always capture the breadth of authentic experiences of the caregiver. Online social media platforms provide a space for individuals to share their experiences and obtain advice toward caring for those with ADRD. This study leverages Reddit, a platform frequented by caregivers seeking advice for caring for a family member with advice for ADRD. To identify the topics of discussion or advice that most caregivers seek and sought after, we employed structured topic modeling techniques such as BERTopic to analyze the content of these posts and use an intertopic distance map to discern the variation in themes across different Reddit categories. In addition, we analyze the sentiment of the Reddit postings using Valence Aware Dictionary and Sentiment Reasoner to deduce the degree of negative, positive, and neutral sentiment of the discussion posts. Our findings reveal that the topics that caregivers most frequently discuss and seek advice for were related to caregiver stories, community support, and concerns ADRD. Specifically, we aimed to reproduce an organic Reddit search of caregiving of abuse on family member, financial struggles, symptoms of hallucinations, and repetition in ADRD family members. These results underscore the importance of online communities for gaining a comprehensive understanding of the multifaceted experiences and challenges faced by ADRD caregivers.

Keywords: Alzheimer's disease and related dementias; Alzheimer's; Dementia; Caregiver; Reddit; Social media; Natural language processing; Sentiment analysis; Topic analysis; BERTopic

***Corresponding author:**
Yulin Hswen
(yulin.hswen@ucsf.edu)

Citation: Hswen Y, Xiong J, Hurley M, Nguyen TT. Experiences of Alzheimer's disease and related dementia family caregivers on Reddit communities: A topic modeling and sentiment analysis. *Artif Intell Health*. 2024;1(3):127-135. doi: 10.36922/aih.3075

Received: March 4, 2024

Accepted: June 7, 2024

Published Online: July 30, 2024

1. Introduction

Alzheimer's disease and related dementias (ADRD) are a group of neurodegenerative disorders that are characterized by impaired thinking and independence.[1] In 2023, an estimated 6.7 million Americans aged 65 and older were living with ADRD.[2] ADRD features a gradual development, and as the disease progresses, behavioral and cognitive functions decline, affecting memory, comprehension, language, attention, reasoning, and judgment.[3] People living with ADRD experience a range of symptoms that interfere with daily life. Memory loss, confusion, and poor judgment are often the first signs that appear in the affected individuals.[4] Overtime, as the disease progresses from mild to moderate and from moderate to severe, additional compounding symptoms arise such as withdrawal from social activities, shortened attention span, difficulty carrying out familiar tasks, impulsive behavior, changes in sleeping patterns, difficulty recognizing friends and family, and general physical decline.[4] Intensive supervision and care become necessary for patients. Managing increasing symptoms and needs can be difficult and highly stressful for patients as well as their families and caregivers.

Depending on the patient's circumstances and needs, caregiving settings available to ADRD patients vary widely, with options for in-home care, adult care centers, long-term care facilities, and hospice.[5] Many patients transition between care settings, posing a challenge for families, caregivers, and providers to ensure coordinated, continuous care.[2] In 2020, the total health-care costs for treatment of ADRD was estimated at $305 billion, most of which is attributed to the cost of care.[6] However, indirect costs of care such as quality of living and informal caregiving are likely underestimated.[6] Patients generally receive increasing care from family members and other unpaid caregivers as their disease progresses[2]. It has been estimated that 75% of caregiving given to a patient with Alzheimer's disease is informal care provided by the patient's family and friends[6]. In 2022, more than 11 million family members and unpaid caregivers provided an estimated 18 billion hours of care to people with ADRD.[2] On top of the emotional burden of caring for a family member with declining cognitive ability, navigating the care system is time-consuming and costly, extending to family caregivers' increased risk for negative mental and physical health outcomes.[2]

Family caregivers play a critical role in the quality of life of ADRD patients, yet their mental health and support circles have been understudied. Research on the experiences and needs of caregivers would not only benefit other caregivers in similar situations but also provide insight for health-care providers and researchers to understand and respond to the needs of these individuals.[7] Existing studies have gathered data on caregivers' well-being through methods such as questionnaires and interviews,[8,9] and recent trends saw the emerging data analysis of caregiver experiences from various social media platforms – an investigation strategy that could avoid response bias.[10,11] Social media platforms allow for the collection of large amounts of content in the form of users' health-care knowledge, experiences, symptoms, products, doctors, and medicines for research purposes[11] and offer insights from otherwise hard-to-reach populations.[12]

For instance, empirical research is increasingly turning to Reddit as a source of information due to its large user base and highly specialized subreddit communities.[12] Compared to other social media platforms, Reddit has a lower noise-to-information ratio, which makes it ideal for people who are caring for family members with ADRD. Recent studies have begun to analyze Reddit posts to research the experiences and challenges of caregivers of patients with Alzheimer's disease and other forms of dementia.[7,13] This paper builds on existing studies and focuses on the important topics of caregivers of family members with Alzheimer's disease. Data are analyzed with topic modeling and sentiment analysis. The analysis reveals the major concerns of Alzheimer's patient's family caregivers, offering insight into future care and assistance to this population.

2. Data and methods

2.1. Data source and cleaning

To mimic an organic search on Reddit, "caregiving for a family member with dementia" was entered into the search box and searched on Reddit. Using the Reddit API, the most relevant posts were extracted. Reddit API allows for the return of information within a post, including user ID, post title, post content, post time, and all comments. In total, 1151 replies were collected.

Text from posts and comments were preprocessed before topic analysis. The goal of preprocessing is to keep words that carry essential meaning. Thus, we remove stopwords using NLTK stopwords dictionary.[14] Stopwords such as "a" and "the" are words that do not convey important information and add little to the comprehension of the text.[15] This step denoises the text input. Next, all words were lemmatized using SpaCy.[16] In this step, all words were swapped with their lemma, so that only the content of the words remained. This step ensures that words in different forms are counted as one in the topic analysis.

2.2. Topic modeling

To understand the top topics within the Reddit comments, we performed topic modeling, a machine learning-based

classification method for texts.[17] In this study, we used BERTopic, a sentence-transformers model, for extracting embedded document. Compared to previous methods such as Latent Dirichlet Allocation (LDA) modeling,[18] BERTopic incorporates the semantic context of words and further fine-grained the method by considering the varying word semantic distance distributions.[19] Similar to the user interface of other topic models, it outputs topic assignment for each comment, as well as the top words of each topic. The top words help us interpret the topics of the comments, while topic assignment lets us see how popular each topic is, and it can also be used in the subsequent sentiment analysis.

Another difference between BERTopic and LDA modeling is that BERTopic determines the number of topics by the text, while LDA relies on a user-defined number of topics.[20,21] Using BERTopic, we generated an intertopic distance map to determine the distance (difference) between the topics. An intertopic distance map represents each topic as a circle on Cartesian plane, whose coordinates represent semantic distance. If circles do not overlap, it is considered that the topics are well separated. If not, the topic model will be refitted with an adequately smaller topic number, and the intertopic distance map will be plotted again to see if the topics are well separated. The "step-size" of each refitting can vary depending on prior knowledge on the dataset. For example, in the case where no more than 20 topics are expected in the text, and BERTopic model identifies more than 100 topics, the "step size" can be 5 – 10 less topics for next refitting, until topic separation appears, or that number of topics is reduced to 20. After that, the "step size" can be 1 less topic for each refitting.

2.3. Sentiment analysis

To understand the sentiment that a comment carries, we performed sentiment analysis, which quantifies positive and negative sentiment. We adopted the most widely used sentiment analysis, Valence Aware Dictionary for Sentiment Reasoning (VADER), for our purpose in this study.[20] VADER is a rule-based model that summarizes lexical, grammatical, and syntactical features of text and quantifies the tone of sentiment into scores.[20] Compound VADER scores are normalized from the raw VADER scores and span from −1 to 1, with a negative score representing negative sentiment, and vice versa. We followed the rule of thumb in VADER sentiment analysis and identified those with compound VADER scores <-0.05 as negative comments, −0.05 to 0.05 as neutral, and those with compound VADER scores >0.05 to be positive comments.

3. Results

3.1. BERTTopic modeling output

A total of 1151 comments were collected from 15 Reddit posts from our search results.[19] Using BertTopic topic modeling and manual topic refinement, we categorized the comments into six topics and provide example comments for each topic in Table 1. Topic 0 was identified as "sharing caregiver stories," topic 1 as "appreciation of online community," topic 2 as "concerns of abuse of ADRD family member," topic 3 as "financial struggles of caregivers," topic 4 as "early symptoms of ADRD of family member," and topic 5 as "symptoms of ADRD." As seen in Table 1, the topic having the greatest proportion of discussions was topic 0 (n = 926), followed by topic 1 (n = 126), topic 2 (n = 33), topic 3 (n = 31), topic 4 (n = 22), and topic 5 (n = 13).

3.2. VADER (sentiment analysis) results

We used VADER to analyze the sentiment of the comments under each topic. Figure 1 describes the average VADER sentiment score of the retrieved posts' texts for each topic. In Figure 1, the x-axis corresponds to the VADER compound score that ranges from −1 to 1, where $x<-0.05$ represents negative sentiment, $-0.05<x<0.05$ represents neutral sentiment, and $x>0.05$ represents positive sentiment. As described by the histogram bars in Figure 1, topic 3 is skewed to the right indicating more positive sentiment, while topics 1 and 3 are skewed to the left indicating more negative sentiment. Figure 2 provides a direct comparison of comment sentiment proportions. Topic 0 had relatively equal proportions of positive and negative sentiment, whereas topic 5 had the most proportion of neutral sentiment and topic 3 had the highest proportion of positive posts.

The top words in each topic are displayed in Table 2. Topic 0 was the largest topic of posts and manually labeled as "shared stories by caregivers." This topic included stories that ADRD caregivers shared with other ADRD caregiving users on Reddit. Comments included personalized experiences of their family member having ADRD symptoms, describing in detail specific cases. Top keywords included specific family members, such as "mom" and "dad." As shown in Table 1 and Figure 1, 44.8% of the posts were negative and 53.2% of posts had a positive sentiment.

Topic 1 was manually labeled as "appreciation of online community." This topic included comments in which caregivers shared gratitude and thanks with other Reddit users, showcasing the benefit of these online communities. The top five keywords in Topic 1 were "thank," "sorry," "much," "go," and "share." As shown in Table 1 and Figure 1,

Table 1. BERTopics, keywords, and sentiment from posts on caregiving for family members with ADRD on Reddit

Topic	Number of comments	Top 5 keywords	Topic interpreted	# Negative comments (%)	# Positive comments (%)	Example text
0	926	Get, go, time, thing, know	Sharing of caregiver stories	415 (44.8)	462 (49.9)	"*First, it was little things – but very noticeable – such as words. A few months later, it was people; she would confuse people from her past and the present. Then, she started hallucinating and making up stories about people that didn't exist doing things with her she couldn't have possibly done (like going to places that no longer existed in my hometown or visiting friends who were dead). Next, she completely forgot about my grandpa (who'd been dead for ten years at this point, and to whom she'd been married for over 50), and after that, she also forgot about her firstborn son, who had died 8 years before.*"
1	126	Thank, sorry, much, go, share	Appreciation of online community	31 (24.6)	67 (53.2)	"*I appreciate you sharing this. Lots of us are in the same boat with you.*"
2	33	Abuse, forgiveness, peace, abuser, heart	Concern of abuse on ADRD family member	12 (36.4)	19 (57.6)	"*Not if they don't want to be. I feel sad for the person he is now because dementia really does change who you are, but I do not fault the family. They still see their abuser when they look at him since he still has the same face, and it doesn't make sense to revictimize them so he can have extra company.*"
3	31	Medicaid, state, care, asset, qualify	Financial struggles of caregivers	3 (9.7)	26 (83.9)	"*How do you afford such care in America? Insurance only covers 28 days (classified as rehab) and the rest is out of pocket. And it's very expensive. Soon the accounts are empty and they go on Medicaid, right? Eventually drying up all assets.*"
4	22	Hallucination, cat, see, real, brain	Early signs of ADRD in family member	6 (27.3)	12 (54.5)	"*The first thing we noticed was paranoia. It started as vague and even somewhat plausible and over time just got more and more extreme. She lives in assisted living now and pretty much everyone who visits her has said that they're taking up dangerous extreme sports because that is no way for a person to die.*"
5	13	Repeat, story, plan, get, weekend	Symptoms of ADRD in family members	3 (23.1)	3 (23.1)	Two minutes later: "*Have you got plans for the weekend?*"

Abbreviation: ADRD: Alzheimer's disease and related dementias.

24.6% of the posts had a negative sentiment and 53.2% had a positive sentiment.

Topic 2 was identified by BERTopic and manually labeled as "concerns of abuse for ADRD family members." This topic included posts in which caregivers described fears of ADRD family members who were patient abuse at care facilities. The top five keywords in Topic 1 were "abuse," "forgiveness," "peace," "abuser," and "heart." As shown in Table 1 and Figure 1, 36.4% of the posts in Topic 2 had a negative sentiment and 57.6% had a positive sentiment. Despite Topic 2 having a higher proportion of positive comments than negative comments, the negative posts had the highest mean negative VADER scores among all the topics (Figure 1).

Topic 3 was identified by BertTopic and manually labeled as "financial struggles of caregivers." This topic included comments related to the financial situations and complexities faced in taking care of a family with ADRD by caregivers. The top five keywords in Topic 3 were "medicaid," "state," "care," "asset," and "qualify." As shown in Table 1 and Figure 1, 9.7% of the comments in Topic 3 had a negative sentiment and 83.9% had a positive sentiment. Topic 3 had the highest proportion of positive comments as well as the highest mean positive VADER scores among all the topics (Figure 1).

Topic 4 was identified by BERTopic and manually labeled as "early signs of ADRD in family members." This topic included posts in which caregivers discussed their family members' early symptoms of ADRD. The top five keywords in Topic 4 were "hallucination," "cat," "see," "real," and "brain." It appears that the first signs of ADRD in family members started to surface when they began to hallucinate and were neglecting their pet cats. As shown

Table 2. Top words by topic

Topic 0		Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
Count	926	Count	126	Count	33	Count	31	Count	22	Count	13
Keyword	*Frequency*	*Keyword*	*Frequency*	*Keyword*	*Frequency*	*Keyword*	*Frequency*	*Keyword*	*Frequency*	*Keyword*	*Frequency*
Get	0.038	Thank	0.425	Abuse	0.207	Medicaid	0.176	Hallucination	0.227	Repeat	0.289
Go	0.033	Sorry	0.190	Forgiveness	0.173	State	0.107	Cat	0.167	Story	0.286
Time	0.031	Much	0.122	Peace	0.139	Care	0.082	See	0.090	Plan	0.146
Thing	0.027	Go	0.095	Abuser	0.112	Asset	0.072	Hallucinate	0.075	Get	0.123
Know	0.025	Share	0.093	Heart	0.103	Qualify	0.069	Real	0.074	Weekend	0.119
Start	0.024	Lovely	0.090	Cut	0.095	Cost	0.063	Brain	0.071	Tell	0.118
Year	0.024	Say	0.082	Family	0.079	Pay	0.060	Look	0.055	Minute	0.115
Mom	0.023	Person	0.064	Deserve	0.079	Insurance	0.055	Even	0.052	Time	0.094
Say	0.022	Funny	0.063	Forgive	0.067	Social	0.053	Go	0.047	Question	0.085
Dad	0.021	Man	0.062	Tie	0.065	Cover	0.047	Dishwasher	0.047	Familiar	0.081

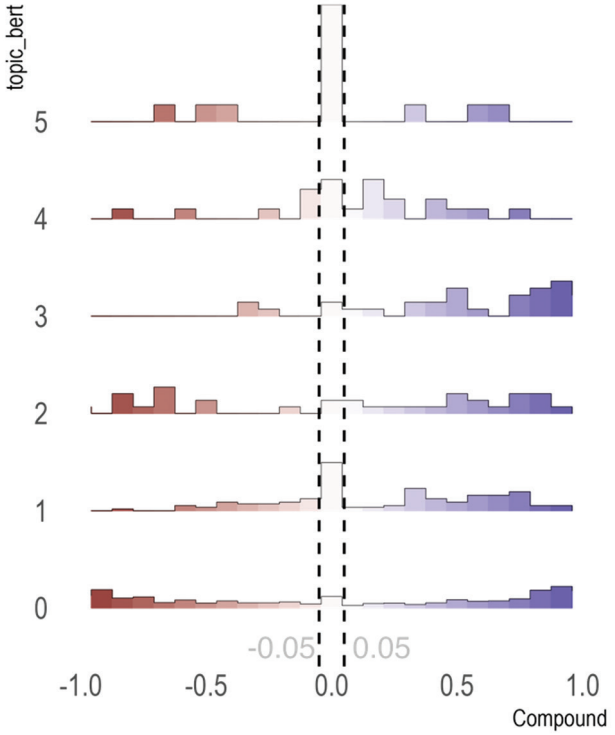


Figure 1. Sentiment distribution of topics. Source: By authors

in Table 1 and Figure 1, 27.3% of the posts had a negative sentiment and 54.5% had a positive sentiment.

Topic 5 was identified by BERTopic and manually labeled as "symptoms of ADRD in family members." This topic included posts in which caregivers discussed their family members' most frequent symptoms of ADRD. The top five keywords in Topic 5 were "repeat," "story," "plan," "get," and "weekend." As shown in Table 1 and Figure 1,

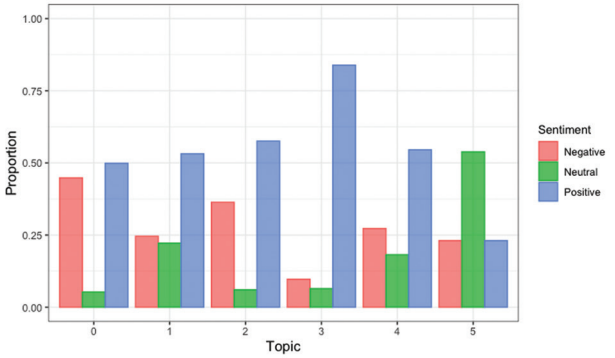


Figure 2. Proportion of negative, neutral, and positive sentiment across topics. Source: By authors

23.1% of the posts had a negative sentiment and 23.1% had a positive sentiment. For instance, family members with ADRD have symptoms of repeating stories and questions and forgetting plans for the weekend.

3.3. Intertopic distance map

Figure 3 displays the intertopic distance map showing a visualization of the topics, with the area of the topic circles proportional to the number of words that belong to each topic and the distance between the topics representing the degree of difference between each topic. From Figure 3, we can see that all topics are well separated. Among the topics, smaller clusters can be observed, where topics 0, 2, and 5 are clustered, while topics 1, 3, and 4 are clustered to each other. The clustering and proximity of topics to one another indicate that the texts in these topic clusters were related to one other semantically. It also follows that topics that are further distanced share less similarity semantically.

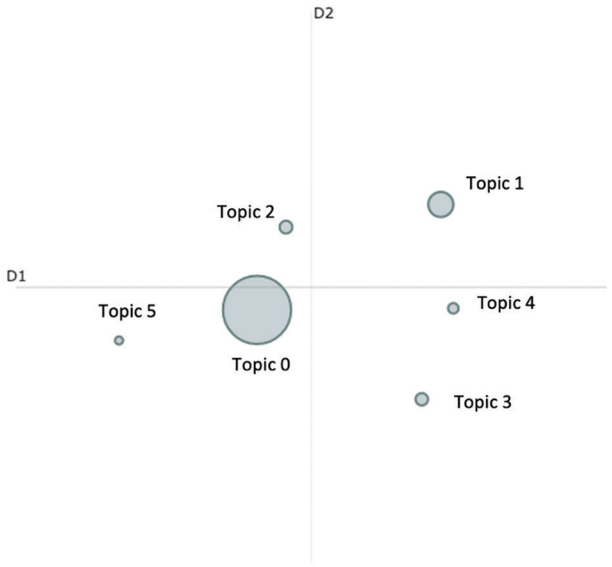


Figure 3. Intertopic distance map for all topics. Source: By authors

4. Discussion

The objective of this study was to determine how Reddit provides online support to caregivers of ADRD. By performing an organic search and collecting the top Reddit postings, we revealed the most prominent topics of discussion among family caregivers of ADRD patients. These results bring to light to perspectives and experiences of caregivers and determine the sentiment of Reddit community forums. Overall, the sentiment of Reddit posts was neutral but leaned toward a positive skew because of the support that Reddit users who were family caregivers of ADRD patients received.

4.1. Building on previous research

Topic modeling has been frequently used to understand unstructured data such as those from social media. Previous systematic review has evaluated the use of topic modeling on social media and overall, it is effective at understanding themes but has its limitations.[21] We employed standard tokenization and the removal of stop words, which is essential to the analysis of topics. We, however, did not evaluate discussions on social media that were multilingual. It is true that multilingual discussions exist on Reddit but, for the purposes of this study, we only collected texts that were in English. In addition, there are various topic models that can be used, which can dictate the interpretation of the data. However, based on this systematic review, LDA analysis is the most frequently used method as it provides the most robust and clean sectioning of data into topics.[21-24] LDA topic analysis has also been used across various dimensions of health and countries.[23] Furthermore, we employed a sentiment analysis tool VADER as an additional method in combination with topic analysis to further validate how caregivers feel toward caring a family member with dementia.[25-28]

4.2. Burden of ADRD on family members

Our findings point broadly to the substantial impact of ADRD on family caregivers. The physical, emotional, and mental stress of diagnosis extends beyond the patient and to those who care for them as the disease progresses. There is no cure for ADRD, so as patients' cognitive abilities decline, family members become their advocates and care experts. The Reddit comments that we analyzed revealed high levels of distress among family members, which is associated with their increase in responsibilities, difficulty navigating care options, disheartening interactions with symptomatic loved ones, and the strain it puts on other areas of their lives. Some illustrative comments are given in the following:

> *"I feel I have no life and my life revolves around taking care of my dad."* (*Topic 0*)
>
> *"I've worked as a nurse for many, many years, with the last 15 of those being in aged care specialising in care & it's always hardest on the family."* (*Topic 0*)
>
> *"My Mom is the sole caregiver and it has been extremely hard on her as you'd expect. My Dad is in a long term care facility now and she's there with him everyday. I feel like Alzheimers has taken away both of my parents, my Dad but also my Mom."* (*Topic 1*)
>
> *"I love my grandma so much and it's always so disheartening to visit her in the state that she's in."* (*Topic 0*)

4.3. Reddit as a resource and form of support for caregivers

Despite the toll that the disease takes on caregivers, our findings revealed primarily positive comments in the Reddit forum. This points to the sense of community and support that the Reddit online community offers. For instance, when a user posts a comment on Reddit that shares about their caregiving experiences, challenges, or questions, the other users in the forum who have gone through similar caregiving situations respond by sharing their advice, empathy, and condolences. Several illustrative comments are presented below:

> "*I'm so sorry you are so overwhelmed. I hope you have someone there you can talk to about this, and relatives to give you a break from caretaking.*" (*Topic 1*)
>
> "*You have every right to feel overwhelmed to see her deteriorate in front of you and want the best*

care for her. It's a total disruption to your life with added responsibilities to have to suddenly care for someone with." (*Topic 1*)

"*Don't be *too* hard on yourself if possible. And if you ever want to talk with (or at) someone that might be able to relate a bit you (and anyone else dealing with this kind of thing) are absolutely, seriously, earnestly, always welcome to message me.*" (*Topic 1*)

"*Good luck. Please remember to forgive yourself for any frustration you feel. It is natural, normal, and would be weird for you not to have any.*" (*Topic 0*)

As family members manage numerous symptoms that progress over time and vary across ADRD patients, Reddit serves as a forum for caregivers to share tips and recommendations that they have learned. Navigating the care systems, symptoms, and treatment for ADRD is complex and daunting. For caregivers researching their loved ones' symptoms and care situations, Reddit serves as a wealth of knowledge curated by those with first-hand experience. Our findings point to Reddit as a resource to medical for patient care management, supplementary to the treatment and advice that medical practitioners provide. Several illustrative comments are shown in the following:

"*My wife and I have been going through it with her dad for several years now. He has and is just now starting to have issues with urine control. We believe it's a side effect from one of his meds. I advise you check her meds and make adjustments if you can.*" (*Topic 0*)

"*We are looking for a memory care facility rather than a nursing home and I recommend you research the same.*" (*Topic 0*)

"*After she started becoming more physical with me we decided to put her in a secured facility (so she wouldn't wander off) It was great! They had activities and a routine for her which really eased my guilt for placing her there. Maybe that's something you should consider.*" (*Topic 1*)

This study underlines the importance of Reddit as a resource for caregivers who may be looking for a forum for managing care, sharing experiences, and finding support.

4.4. Reddit as a resource for practitioners, researchers, and health-care organizations

While it is evident that Reddit forums act as a form of support and guidance for caregivers, our findings can also be applied in professional health-care settings. The sentiment analysis of recurring topics in Reddit threads related to Alzheimer's disease informs health-care providers of some of the most common symptoms that caregivers struggle to manage. Reddit's candid format brings unbiased perspectives and insight into the way that resources, tools, and educational services should be diverted in health-care settings for caregivers and patients. A representative feedback below illustrates this:

"*The hardest part for me has been dealing with the anger, paranoia, and confabulation that my mom presents with.*" (*Topic 1*)

5. Limitations

A limitation of this study is its reliance on Reddit for data, which, despite providing a rich dataset of caregiver experiences, might not capture the full spectrum of perspectives available on the array of social media platforms. Incorporating data from other platforms such as Facebook groups, Twitter, and other online forums could offer greater insight. Different platforms often cater to diverse demographics and feature varied caregiving experiences, which could enrich the overall analysis by presenting a broader range of insights. For this study, we focused on using Reddit data because of its specific forums for ADRD caregivers.

A second limitation of the current study is its cross-sectional design, which does not track changes in discussions and sentiments of ADRD caregivers over time. With the anonymous nature of Reddit, we are unable to track users longitudinally. A longitudinal study would provide valuable insights into how caregiving challenges and needs evolve, reflecting the dynamic nature of caregiver experiences as the disease progresses or in response to changes in social and health care that ADRD family members receive. Such an approach would allow for the identification of trends and shifts in caregiver concerns, offering a deeper understanding of the temporal dynamics at play. However, the anonymity of this social media data is also critical to capturing organic and honest personal caregiving experiences. Research should continue to prioritize privacy protections and confidentiality for this social media data.

6. Future directions

In future studies, new artificial intelligence (AI)-based approaches for the analysis of caregiver discussions should be explored. Specifically, transformer-based models, such as GPT-3 and its iterations, could provide a more nuanced understanding of context, sentiment, and emotional undertones. The next step in subsequent studies should be to test these new generative AI tools to determine whether they can match or surpass the performance of currently validated natural language processing techniques. This would enhance the field and also ensure that we evaluate

what is the most effective AI tool in understanding unstructured human communications.

This study focused on the top posting on Reddit about caregiving for family members with ADRD and did not explore the variability of caregiving experiences and sentiments across different cultures, geographic locations, and demographic groups, including age, gender, and socio-economic status. The challenge with using Reddit is the preservation of users' anonymity. On top of that, subanalyses specific to different demographics may illuminate the unique challenges and needs faced by caregivers from diverse backgrounds. Efforts to understand these differences are crucial for the development of targeted support and resources that are tailored to meet the specific needs of ADRD caregivers. Identifying distinct social factors among ADRD patient's caregivers can help devise support that is inclusive and responsive to the diverse realities of caregivers worldwide.

Finally, comparing insights from social media data of ADRD caregivers with data from electronic health records and clinical notes concerning ADRD may provide further validation of these online data sources. This integrative approach would allow researchers to identify potential correlations between the topics and sentiments expressed in social media discussions and actual clinical outcomes. By mapping these discussions to specific stages of ADRD, it may be possible to discern patterns and trends that inform more effective caregiver support strategies. This method could prove invaluable in enhancing our understanding of how caregiver experiences and needs evolve in response to the progression of the disease and the efficacy of interventions.

7. Conclusion

This study uses sentiment and topic analysis, to disentangle posts on Reddit on how caregivers or patients themselves are self-managing care. Our findings can also be useful for bridging the gap between theoretical insights derived from social media discussions and actionable recommendations. Health-care providers can use this information to translate the nuanced understanding of caregiver experiences from these social media sources into improved support strategies and patient care interventions, ultimately benefiting both caregivers and patients by ensuring that the insights gained are effectively applied in health-care settings.

Acknowledgments

None.

Funding

Research reported in this publication was supported by the National Institute on Minority Health and Health Disparities (R00MD012615 [TTN], R01MD015716 [TTN]). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Yulin Hswen, Jiangmei Xiong, Thu T. Nguyen
Investigation: Yulin Hswen, Jiangmei Xiong, Thu T. Nguyen
Methodology: Yulin Hswen, Jiangmei Xiong
Writing – original draft: Yulin Hswen, Jiangmei Xiong, Margaret Hurley
Writing – review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data will be available upon reasonable request.

References

1. Breijyeh Z, Karaman R. Comprehensive review on Alzheimer's disease: Causes and treatment. *Molecules*. 2020;25(24):5789.

   doi: 10.3390/molecules25245789

2. 2023 Alzheimer's disease facts and figures. *Alzheimers Dement*. 2023;19(4):1598-1695.

   doi: 10.1002/alz.13016

3. Kumar A, Sidhu J, Goyal A, Tsao JW. Alzheimer disease. In: *StatPearls*. Treasure Island, FL: StatPearls Publishing; 2024.

4. National Insitute on Aging. *What Are the Signs of Alzheimer's Disease*? Bethesda: National Institute of Health; 2024.

5. Alzheimer's Association. *Care Options*. United States: Alzheimer's Association; 2024.

6. Wong W. Economic burden of Alzheimer disease and managed care considerations. *Am J Manag Care*. 2020;26(8 Suppl):S177-S183.

   doi: 10.37765/ajmc.2020.88482

7. Garg M, Sohn S. CareD: Caregiver's experience with cognitive decline in Reddit posts. *IEEE Int Conf Healthc Inform*. 2023;2023:581-587.

   doi: 10.1109/ichi57859.2023.00104

8. Vu M, Mangal R, Stead T, Lopez-Ortiz C, Ganti L. Impact of Alzheimer's disease on caregivers in the United States. *Health Psychol Res*. 2022;10(3):37454.

   doi: 10.52965/001c.37454

9. Sołtys A, Tyburski E. Predictors of mental health problems in formal and informal caregivers of patients with Alzheimer's disease. *BMC Psychiatry*. 2020;20(1):435.

   doi: 10.1186/s12888-020-02822-7

10. Shoults CC, Rutherford MW, Kemp AS, *et al*. Analysis of caregiver burden expressed in social media discussions. *Int J Environ Res Public Health*. 2023;20(3):1933.

    doi: 10.3390/ijerph20031933

11. Lobo EH, Johnson T, Frølich A, *et al*. Utilization of social media communities for caregiver information support in stroke recovery: An analysis of content and interactions. *PLoS One*. 2022;17(1):e0262919.

    doi: 10.1371/journal.pone.0262919

12. Zapcic I, Fabbri M, Karandikar S. Using Reddit as a source for recruiting participants for in-depth and phenomenological research. *Int J Qual Methods*. 2023;22:16094069231162674.

    doi: 10.1177/16094069231162674

13. Ni C, Malin B, Song L, Jefferson A, Commiskey P, Yin Z. "Rough Day … Need a Hug": Learning challenges and experiences of the Alzheimer's disease and related dementia caregivers on Reddit. *Proc Int AAAI Conf Web Soc Media*. 2022;16(1):711-722.

    doi: 10.1609/icwsm.v16i1.19328

14. Bird S, Klein, E, Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. United States: O'Reilly Media, Inc.; 2009.

15. Méndez JR, Iglesias EL, Fdez-Riverola F, Díaz F, Corchado JM. *Tokenising, Stemming and Stopword Removal on Anti-spam Filtering Domain*. Heidelberg: Springer Berlin; 2006. p. 449-458.

16. Honnibal M, Montani I, Van Landeghem S, Boyd A. *spaCy: Industrial-strength Natural Language Processing in Python*. United States: Zenodo; 2020.

17. Murakami A, Thompson P, Hunston S, Vajn D. What is this corpus about?': Using topic modelling to explore a specialised corpus. *Corpora*. 2017;12:243-277.

    doi: 10.3366/cor.2017.0118

18. Blei DM. Probabilistic topic models. *Commun ACM*. 2012;55(4):77-84.

    doi: 10.1145/2133806.2133826

19. Grootendorst MR. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv*. 2022.

    doi: 10.48550/arXiv.2203.05794

20. Hutto CJ, Gilbert E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In: *Proceedings of the 8th International Conference on Weblogs and Social Media*. ICWSM; 2015.

21. Laureate CDP, Buntine W, Linger H. A systematic review of the use of topic models for short text social media analysis. *Artif Intell Rev*. 2023;56:14223-14255.

    doi: 10.1007/s10462-023-10471-x

22. Kherwa P, Bansal P. Topic modeling: A comprehensive review. *EAI Endorsed Trans Scalable Inf Syst*. 2019;7(24):e2.

    doi: 10.4108/eai.13-7-2018.159623

23. Ramamoorthy T, Kulothungan V, Mappillairaju B. Topic modeling and social network analysis approach to explore diabetes discourse on Twitter in India. *Front Artif Intell*. 2024;7:1329185.

    doi: 10.3389/frai.2024.1329185

24. Yue L, Chen W, Li X, Zuo W, Yin M. A survey of sentiment analysis in social media. *Knowl Inf Syst*. 2019;60:617-663.

    doi: 10.1007/s10115-018-1236-4

25. Rodríguez-Ibánez M, Casánez-Ventura A, Castejón-Mateos F, Cuenca-Jiménez, PM. A review on sentiment analysis from social media platforms. *Expert Syst Appl*. 2023;223:119862.

    doi: 10.1016/j.eswa.2023.119862

26. Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. *Artif Intell Rev*. 2022;55:5731-5780.

    doi: 10.1007/s10462-022-10144-1

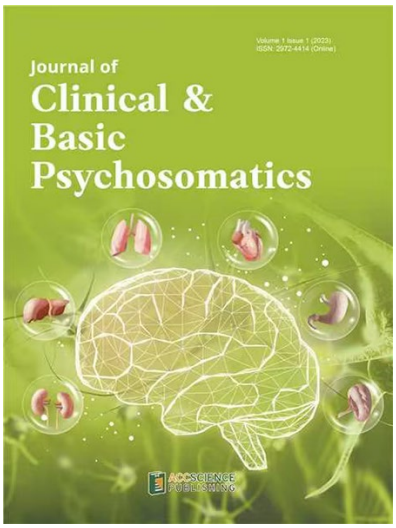
27. Caschera MC, Ferri F, Grifoni P. Sentiment Analysis from Textual to Multimodal Features in Digital Environments. In: *Proceedings of the 8th International Conference on Management of Digital EcoSystems* (*MEDES*). New York, USA: Association for Computing Machinery; 2016. p. 137-144.

    doi: 10.1145/3012071.3012089

28. Xu QA, Chang V, Jayne C. A systematic review of social media-based Sentiment analysis: Emerging trends and challenges. *Decis Anal J*. 2022;3:100073.

    doi: 10.1016/j.dajour.2022.100073

Our Journals



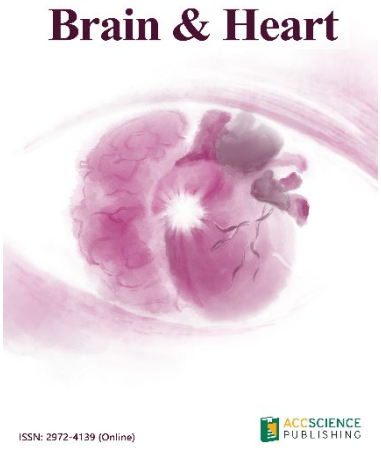
Journal of Clinical and Basic Psychosomatics (JCBP) is a quarterly journal focusing on clinical and basic research on symptoms, assessment, treatment, management, and the mechanism of psychosomatic disorders. *Journal of Clinical and Basic Psychosomatics* covers subject areas, including but not limited 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