SPAST Open Access Research Article

Machine Learning in Healthcare Predictive Analytics: A Comprehensive Review

Ajay Kumar[1], Midhun Chakkaravarthy[2], and Shashi Kant Gupta[3]

[1,2,3]Lincoln University College, Malaysia

Abstract

The purpose of writing this article is to use machine learning approaches to study human predicting diseases. After reviewing previous research, this paper separates the various machine learning algorithms for different diseases based on the quality of the data and their results. In order to improve healthcare decision-making, machine learning-based predictive models and the identification of research gaps are also included. Various machine learning methods, such as drug response prediction, have been applied to individualized patient care. For time-series, imaging, genomic, structured, and unstructured data, the bioinformatics division explains the importance of data sources. Limitations and difficulties in healthcare forecasting are examined. Emerging tendencies and future orientations are also concluded.

keywords: machine learning; predicting diseases; healthcare decision-making; patient care.

Introduction

Everybody in the world is aware of the disorders that are diagnosed on a regular basis. For the sake of the human race's well-being, the World Health Organization (WHO) of the United Nations establishes Sustainable Development Goals. In order to rank the nation from m̈ost healthyẗo l̈east healthy, a number of organizations take into account the important aspect of different health indicators. Key characteristics include life expectancy, infant mortality rate, health spending per capita, access to healthcare, disease prevalence, lifestyle, environmental, mental health, and social determinants of health. These elements are brought on by human health problems that could be regarded as essential to overall health and wellbeing. According to statistics [1] from the social media magazine V̈isual Capitalist, the worldś poorest and healthiest nations are ranked in the figure 1. According to this health ranking, there are many problems in nations like India and Africa where the welfare of the human race is not given priority. However, there are a few so-called industrialized countries, like Japan and Singapore, that continually prioritize health.

Author Email: pdf.ajaykumar@lincoln.edu.my

The data type gathered, early disease prediction, patient risk stratification, etc., utilizing machine learning algorithms are the main topics of this article. In the field, machine learning algorithms are thought to play the most significant role in enhancing healthcare decision-making. The operational effectiveness of Electronic Health Records (EHR) depends on the vast amounts of patient data generated by healthcare institutions. A subset of artificial intelligence called machine learning algorithms can recognize patterns in data and forecast outcomes from intricate datasets. This survey examines the methods, uses, and difficulties of machine learning (ML) in healthcare predictive analytics.

There are five sections in this study. The inhabitants' basic awareness of health and well-being is introduced in the first section. A systematic literature framework and a review of the literature survey and research gaps are covered in the second section. Part three provides an explanation of machine learning techniques. The outcome and debate are covered in the fourth section. Future directions were concluded in the final section.

Related works & Research gaps

The literature review conducted by current researchers is covered in this part. They conclude the results and provide an explanation of the approaches used on the healthcare datasets. The importance of machine learning in healthcare analytics is examined by the author [2], who focuses on supervised learning for predictions and unsupervised learning for pattern recognition. Through improved data analysis and prediction, these methods improve clinical decision support and public health interventions for high-risk populations.

Tulsani *et.al.* [3] investigates Explainable AI (XAI) in medical imaging, which improves AI interpretability and improves patient care and diagnostic accuracy. It draws attention to developments in XAI, moral dilemmas, and difficulties with black-box AI models. The study addresses issues including data quality, ethics, and human-AI interaction while discussing interpretable machine learning approaches, such as deep learning, post hoc methods, and linear models. Nasr *et.al.* [4] examines smart healthcare frameworks, focusing on sensor integration, security, and machine learning in remote monitoring. It emphasizes modular architectures, real-time response, and cloud integration. Key methods include smart sensors for health vitals, predictive analytics, activity tracking with Bluetooth, and blood pressure measurement using regression techniques.

Yulia *et.al.* [5] explores STD classification using data mining. K-NN achieved 90% accuracy but faced misclassification issues. Future work focuses on optimizing K-NN for better results.Leukemia diagnosis problems, such as data limits and manual errors [6] are covered. Accuracy is improved via segmentation techniques, deep learning, and machine learning. With an emphasis on patient risk models and AI problems, Kusuma *et.al.* [7] investigates the use of big data

Author Email: pdf.ajaykumar@lincoln.edu.my

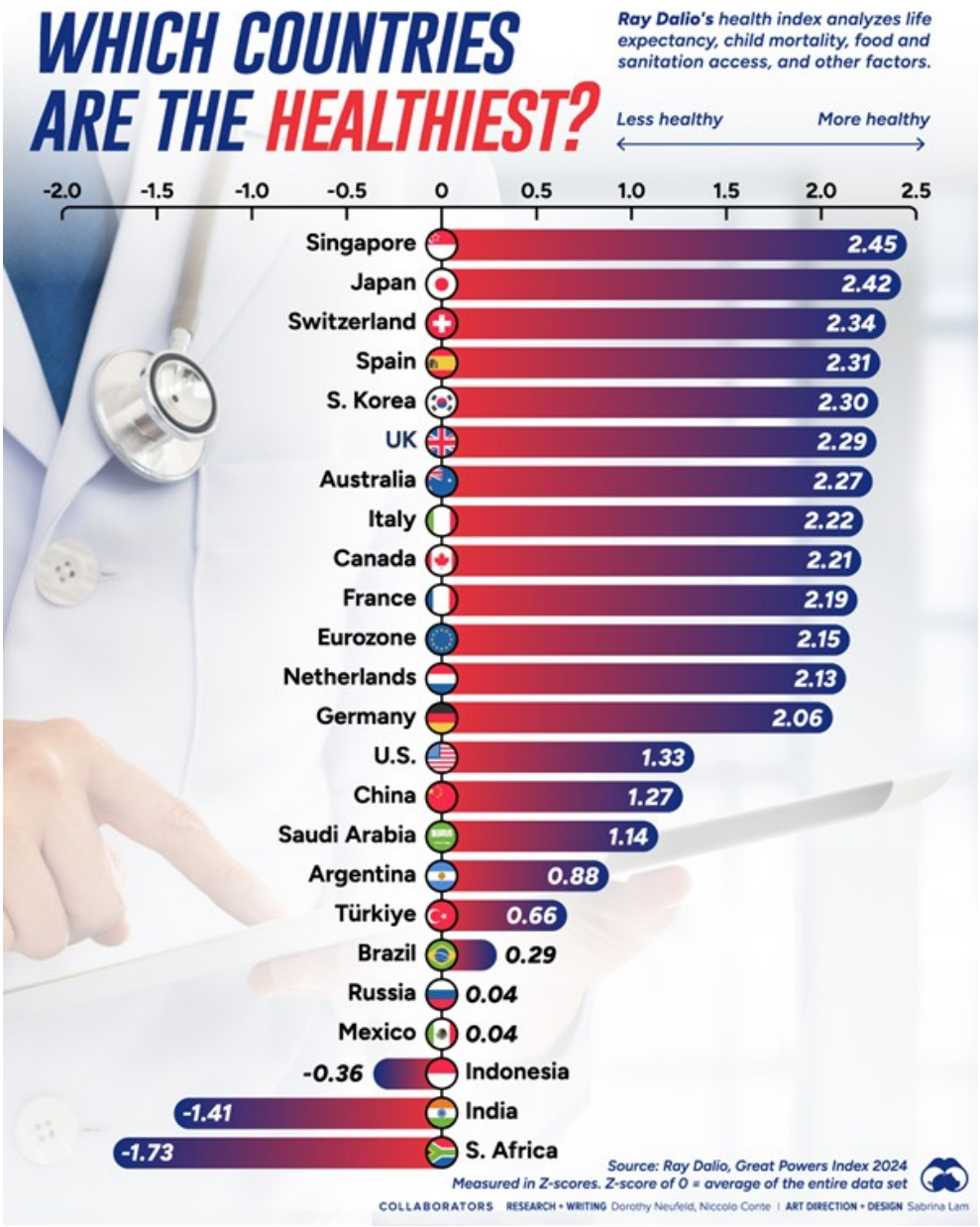

Fig. 1: : Great-Powers-Index-Health_website_Sept24

Author Email: pdf.ajaykumar@lincoln.edu.my

and machine learning in the treatment of cardiac disease. It examines machine learning methods such as support vector machines (SVMs), decision trees, and deep learning, with a focus on early diagnosis and data-driven healthcare enhancements.

Using blood biomarkers, this study [8] creates an ML pipeline to forecast COVID-19 mortality. Normalization and KNN imputation were part of the data preprocessing. Neural networks with XGBoost enhanced feature selection. After evaluating six machine learning models, the neural network's accuracy was 96.53%. Neutrophils, lymphocytes, LDH, hs-CRP, and age are among the important biomarkers that have been found. Single-source data and omitted comorbidity information are among the limitations. In order to balance test cost and accuracy, Ali *et.al.* [9] presents a cost-sensitive feature selection method for CKD diagnosis. Two techniques use the Symmetric Uncertainty and Relief algorithms to maximize feature selection. While lowering expenses, experiments using a CKD dataset demonstrate accuracy that is comparable to baseline models. In order to improve feature interactions, future research will concentrate on improving ensemble ranking. An Enhanced Genetic Algorithm (EAGA) is proposed [10] to optimize feature selection in the categorization of Type 2 diabetes. With an MLP classifier, EAGA achieves 97.96% accuracy by adapting mutation and crossover rates using the Pima Indian Diabetes dataset. It performs better than conventional techniques and holds promise for clinical application as well as wider disease applications. A deep learning method for COVID-19 identification is proposed [11], which uses CGRO for feature selection and ResNet18 for feature extraction. SVM has the highest accuracy when tested on three datasets alongside KNN and ELM. CGRO reduced redundancy and enhanced feature selection. Hybrid algorithms and sophisticated architectures are the main topics of future research.

The knowledge gap identified in the literature review is expressed in the tabular data [Table 1].

Methodology details

The above study enforces to describe the methodologies used in the work aligned for healthcare prediction using machine learning.

AdaBoost [20] is an ensemble learning algorithm that boosts weak classifiers by iteratively adjusting sample weights to improve accuracy. It combines multiple weak learners, focusing on misclassified cases, and makes final predictions through weighted voting. Widely used in various domains, it enhances accuracy but is sensitive to noise and outliers. Particle Swarm Optimization (PSO) [19] is a nature-inspired algorithm where particles explore a search space, adjusting positions based on personal and global bests. It balances exploration and exploitation, making it efficient for optimization tasks in various fields. Federated learning [17] enables secure, decentralized training of AI models in healthcare by sharing model updates instead of patient data, ensuring privacy while improving

Author Email: pdf.ajaykumar@lincoln.edu.my

Tab. 1: Knowledge gaps findings

References [ ]	**Method used**	**Challenges**
Rana *et.al.*, [12]	WOA	solving high-dimensional and multimodal problems
Thabtah *et.al.*, [13]	Naive Bayes, Logistic Regression	Time consuming, rare autism datasets
Uddin *et.al.*, [14]	kNN varients	bias towards dependent neighbours, lack of effective distance calcualtion
Hassan *et.al.*, [15]	k-mean, DWT, SOM	High-complexity and low precision
Rastogi *et.al.*, [16]	Xception, VGG19	Limited deep learning model hinders clinical acceptance
Rahman *et.al.*, [17]	Federated Learning, Explainable AI (XAI)	high cost and time consumption in cluster computing
Band *et.al.*, [18]	CNN, SVM	High cost issues affect IoT hardware communication
Singh *et.al.*, [19]	PSO, GA, Mutual Information	Incomplete test data hinder classification accuracy.
Zhao *et.al.*, [20]	SVM, AdaBoost	Noisy EHR Data

diagnostics and treatment. Xception improves efficiency by using depthwise separable convolutions, reducing parameters while maintaining accuracy. VGG, on the other hand, employs stacked 3×3 convolutions, offering simplicity and strong feature extraction capabilities, though with higher computational cost. Discrete Wavelet Transform (DWT) [15] decomposes signals into multi-resolution components using wavelet functions, enabling efficient feature extraction and noise reduction in image and signal processing.

PRISMA Framework

This investigation utilizes the PRISMA framework (Preferred Reporting Item for Systematic Reviews and Meta-Analyses) [21]. Approximately 52 research papers have been examined, and a framework has been deliberately designed with careful consideration of the inclusion and exclusion criteria for the articles. The initial count of articles stands at approximately 52, with the final selection

Author Email: pdf.ajaykumar@lincoln.edu.my

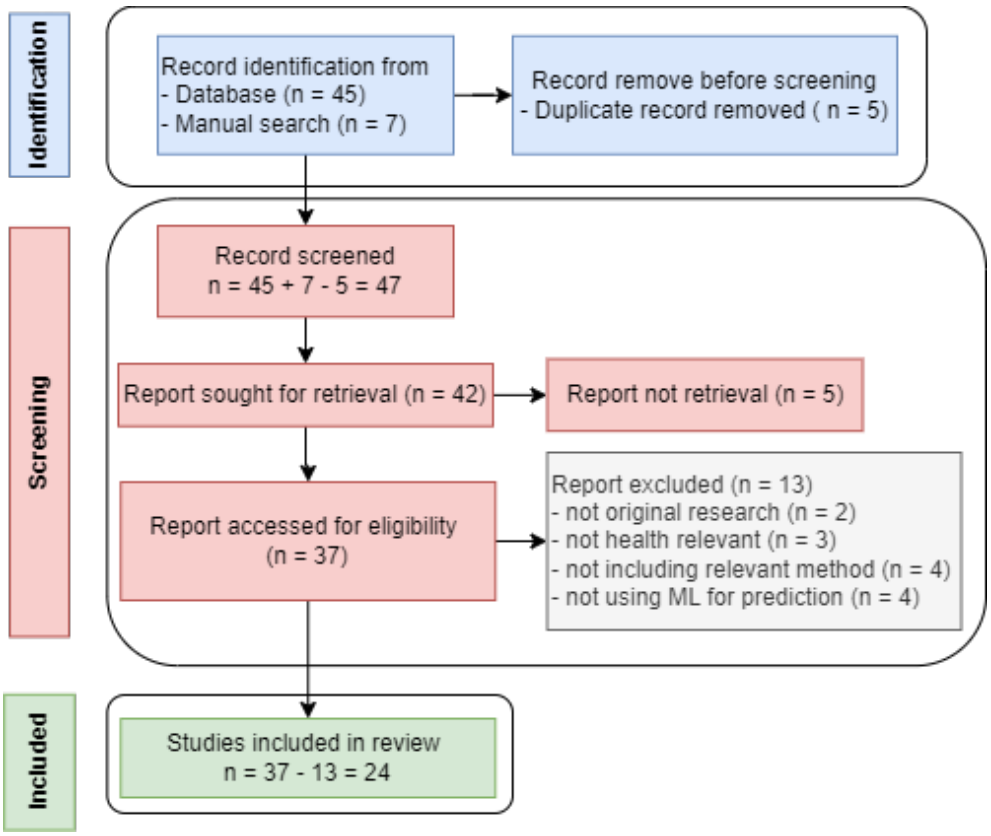

Fig. 2: : PRISMA Framework

resulting in 24, as illustrated in Figure 2.

The fact that a significant portion of the PRISMA framework has been approved for our study direction indicates that the research survey was carried out with care. The three PRISMA sections—identification, screening, and inclusion—are described. In the framework's first phase, Identification, 45 records are selected for consideration, and 7 manual searches are conducted. The same method is used to identify and discard duplicate records (n = 5). The cumulative record is currently 47 in the second section, Screening. Nearly every minute, the report was sought for retrieval (n = 42) rather than non-retrieval (n = 5). Only the eligible records (n = 37) are evaluated; the others (n = 13) are not. Only (n = 24) papers are included for the review in the final component of the framework, Inclusion.

Author Email: pdf.ajaykumar@lincoln.edu.my

Results and Discussions

Investigations on health predictive analysis have primarily concentrated on various health concerns. Research conducted by Zhao *et. al.* [20] and Yuliastuti *et. al.* [5] emphasizes the problem of misclassification within the dataset and the presence of noisy electronic health record data. These studies highlight the significance of non-noisy data, suggesting improved classification outcomes. Recent breakthroughs have resulted in improvements to the genetic algorithm. Mishra *et al.* [10] introduced EAGA, which has enhanced the predictive accuracy for Type 2 Diabetes. A transition from conventional machine learning to advanced machine learning incorporating feature selection has been noted, resulting in a 97.96% improvement in accuracy through the implementation of mutation and crossover rates.

In spite of advancements, there are still certain deficiencies. Thabtah *et.al.* [13], for example, observes that the findings of the Naïve Bayes method were delayed due to the usage of rare autism datasets. Furthermore, multimodal challenges have not received much attention. Conflicting results, like those by Singh *et.al.* [19] and Rana *et.al.* [12], indicate that more research is necessary. Filling up these gaps can improve disease prediction in medical data sets. To enhance the results, future studies could investigate other optimization strategies using primary datasets. Including algorithms inspired by nature may also produce better outcomes.

References

[1] Ray Dalio (2024). The Great Power Index: 2024, Which Countries are the healthiest in 2024, https://www.visualcapitalist.com/which-countries-are-the-healthiest-in-2024/#google_vignette, (Sept 29, 2024)

[2] Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, 30, 100924. https://doi.org/10.1016/j.imu.2022.100924

[3] Tulsani, V., Sahatiya, P., Parmar, J., Parmar, J. (2023). XAI Applications in Medical Imaging: A Survey of Methods and challenges. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(9), 181–186. https://doi.org/10.17762/ijritcc.v11i9.8332

[4] Nasr, M., Islam, M. M., Shehata, S., Karray, F., & Quintana, Y. (2021). Smart healthcare in the age of AI: recent advances, challenges, and future prospects. *IEEE access*, 9, 145248-145270.

[5] Yuliastuti, G. E., Alfiyatin, A. N., Rizki, A. M., Hamdianah, A., Taufiq, H., & Mahmudy, W. F. (2018). Performance Analysis of Data Mining Methods for Sexually Transmitted Disease Classification. *International Journal of Electrical & Computer Engineering* (2088-8708), 8(5).

Author Email: pdf.ajaykumar@lincoln.edu.my

[6] Ghaderzadeh, M., Asadi, F., Hosseini, A., Bashash, D., Abolghasemi, H., & Roshanpour, A. (2021). Machine learning in detection and classification of leukemia using smear blood images: a systematic review. *Scientific Programming*, 2021(1), 9933481.

[7] Kusuma, S., & Udayan, J. D. (2018). Machine learning and deep learning methods in heart disease (HD) research. *International Journal of Pure and Applied Mathematics*, 119(18), 1483-1496.

[8] Karthikeyan, A., Garg, A., Vinod, P. K., & Priyakumar, U. D. (2021). Machine learning based clinical decision support system for early COVID-19 mortality prediction. *Frontiers in Public Health*, 9. https://doi.org/10.3389/fpubh.2021.626697

[9] Ali, S. I., Bilal, H. S. M., Hussain, M., Hussain, J., Satti, F. A., Hussain, M., Park, G. H., Chung, T., & Lee, S. (2020). Ensemble Feature Ranking for Cost-Based Non-Overlapping Groups: A case study of Chronic kidney disease diagnosis in developing countries. *IEEE Access*, 8, 215623–215648. https://doi.org/10.1109/access.2020.3040650

[10] Mishra, S., Tripathy, H. K., Mallick, P. K., Bhoi, A. K., & Barsocchi, P. (2020). EAGA-MLP—An Enhanced and Adaptive Hybrid Classification Model for diabetes diagnosis. *Sensors, 20(14)*, 4036. https://doi.org/10.3390/s20144036

[11] Chattopadhyay, S., Dey, A., Singh, P. K., Geem, Z. W., & Sarkar, R. (2021). COVID-19 Detection by Optimizing Deep Residual Features with Improved Clustering-Based Golden Ratio Optimizer. *Diagnostics, 11(2)*, 315. https://doi.org/10.3390/diagnostics11020315

[12] Rana, N., Latiff, M. S. A., Abdulhamid, S. I. M., & Chiroma, H. (2020). Whale optimization algorithm: a systematic review of contemporary applications, modifications and developments. *Neural Computing and Applications, 32*, 16245-16277.

[13] Thabtah, F. (2018). An accessible and efficient autism screening method for behavioural data and predictive analyses. *Health Informatics Journal, 25(4)*, 1739–1755. https://doi.org/10.1177/1460458218796636

[14] Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports, 12(1)*. https://doi.org/10.1038/s41598-022-10358-x

[15] Hassan, N. S., Abdulazeez, A. M., Zeebaree, D. Q., & Hasan, D. A. (2021). Medical Images Breast Cancer segmentation based on K-Means Clustering Algorithm: a review. *Asian Journal of Research in Computer Science*, 23–38. https://doi.org/10.9734/ajrcos/2021/v9i130212

Author Email: pdf.ajaykumar@lincoln.edu.my

[16] Rastogi, D., Johri, P., Donelli, M., Kumar, L., Bindewari, S., Raghav, A., & Khatri, S. K. (2025). Brain tumor detection and prediction in MRI images utilizing a Fine-Tuned transfer learning model integrated within deep learning frameworks. *Life, 15(3)*, 327. https://doi.org/10.3390/life15030327

[17] Rahman, A., Hossain, M. S., Muhammad, G., Kundu, D., Debnath, T., Rahman, M., Khan, M. S. I., Tiwari, P., & Band, S. S. (2022). Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues. *Cluster Computing, 26(4)*, 2271–2311. https://doi.org/10.1007/s10586-022-03658-4

[18] Band, S. S., Ardabili, S., Yarahmadi, A., Pahlevanzadeh, B., Kiani, A. K., Beheshti, A., Alinejad-Rokny, H., Dehzangi, I., Chang, A., Mosavi, A., & Moslehpour, M. (2022). A Survey on Machine Learning and Internet of Medical Things-Based Approaches for Handling COVID-19: Meta-Analysis. *Frontiers in Public Health, 10*. https://doi.org/10.3389/fpubh.2022.869238

[19] Singh, R. (2018). A Gene Expression Data Classification and Selection Method using Hybrid Meta-heuristic technique. *ICST Transactions on Scalable Information Systems*, 0(0), 159917. https://doi.org/10.4108/eai.13-7-2018.159917

[20] Zhao, J., Henriksson, A., Asker, L., & Boström, H. (2015). Predictive modeling of structured electronic health records for adverse drug event detection. *BMC Medical Informatics and Decision Making, 15(S4)*. https://doi.org/10.1186/1472-6947-15-s4-s1

[21] Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International journal of educational technology in higher education, 20(1)*, 22.

Author Email: pdf.ajaykumar@lincoln.edu.my